DEVELOPMENT AND APPLICATION OF NUCLEAR MAGNETIC RESONANCE TECHNIQUES FOR THE STUDY OF PROTEIN DYNAMICS

by
Bradley J. Harden

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

December 2016

Abstract

The dynamic nature of proteins is a topic of increasing importance in the field of structural biology. Nuclear magnetic resonance (NMR) experiments that characterize protein dynamics offer an unparalleled level of insight into such motional processes. However, a multitude of challenges impede broad application of the technique. The process of assigning each NMR signal to its corresponding atom remains difficult in crowded spectra. Furthermore, lengthy experiments and laborious data analysis can curtail motivations to perform NMR dynamics experiments. In this work, I have developed improvements to two underutilized classes of NMR techniques, and I have applied NMR experiments to study dynamic interactions in an important class of enzymatic systems. I implemented novel covariance NMR processing methods that help minimize artifacts, including a pre-processing spectral derivative that suppresses false positive artifacts stemming from near degeneracy between signals and a post-processing element-wise multiplication step that eliminates coincidental correlations. I applied the improvements to develop a series of novel four-dimensional covariance spectra that streamline the assignment of signals in crowded NMR spectra. I developed a software application to simplify and improve analysis of data acquired with accordion spectroscopy, a technique that can reduce the time required to perform NMR relaxation measurements. I established an approach to fitting accordion NMR data that maximizes accuracy in cases of strong overlap while also ensuring optimal measurement precision. I applied NMR dynamics experiments to study the interaction between a peptidyl carrier protein and its linker regions within a

megadalton non-ribosomal peptide synthetase assembly. I found that interaction between the linker region and the carrier protein core stabilized the latter by a factor of six and modulates its fast, ps-ns time-scale motions. This modulation reveals a dynamic allosteric network that may provide communication between a site of post-translational modification and a linker region to help maneuver the carrier protein during synthesis.

Adviser: Professor Dominique P. Frueh, Ph.D.

Reader: Professor Joel R. Tolman, Ph.D.

Acknowledgements

I would first like to thank Professor Nicholas E. Simpson, Ph.D. and Professor Thomas H. Mareci, Ph.D. who guided my early interest in scientific research and helped me develop a strong foundation in NMR. I would not be where I am today without their support. I would also like to thank Professor David T. Yue, whose passion for science was palpable. Although my time in his lab was brief, his profound dedication and enthusiasm continue to inspire me. He will be greatly missed.

I am indebted to Dr. Ananya Majumdar for his support throughout my time in graduate school. I learned extensively from his carefully crafted lectures and notes, and his pedagogical style continues to influence my own approaches to teaching. Furthermore, his extensive troubleshooting experience saved me many days of frustration with the instruments. I would like to thank the members of my thesis committee for their support and guidance: Professor Caren L. Freel Meyers, Dr. Joel S. Bader and Dr. Joel R. Tolman. I owe a great deal to Dr. Michael T. Morgan and Dr. Alison E. Moreno for their assistance over the past six years. Their meticulously crafted protocols and continued patience with my ignorance guided my development as an experimentalist. I am also grateful to Dr. Andrew C. Goodrich and Dr. Subrata H. Mishra for our extensive discussions and musings on science as well as the stimulating and congenial atmosphere they cultivated. I am eternally grateful to one of my very best friends Dr. Scott R. Nichols, whose broad knowledge and extensive experience in structural biology, biochemistry, microbiology, immunology and more continue to impress me to this day. I came to

the lab with little knowledge of biological phenomena and no practical experience in experimental biology. Everything I know about biology I owe to him. Scott is an exceptional scientist and friend, and I could not have chosen a better person with whom to spend five years sharing an office. I am also thankful to my adviser Professor Dominique P. Frueh. The profound depth of his knowledge in the theory and practice of NMR as well as his encyclopedic knowledge of the literature were instrumental to my success.

I am enormously appreciative to my biological and Thread extended families. Their endless support over the past six years has helped me to persevere in the face of adversity. I also am deeply grateful to Katie Hamblin. Her loyalty and encouragement throughout the many challenges I have faced this year has meant more to me than she knows. Finally, I owe my deepest gratitude to my parents. From an early age, they cultivated my interest in science, math and technology and encouraged me to think critically about all aspects of life. This thesis is dedicated to my father, Roger, who exemplifies unwavering determination in the face of any obstacle and visualized this moment many years before I even conceived of it. And it is dedicated to my mother, Eileen, whose fierce passion, empathy and supreme dedication to her children served as an example that I try to imitate daily. Her spirit lives on in me.

Table of Contents

List of Figures

List of Tables

1 Introduction

Nuclear magnetic resonance (NMR) is an essential tool in studies of protein dynamics, owing in great part to the atomic level description it provides. NMR has been used to probe modulation of protein dynamics during enzymatic reactions[1,2], to relate protein dynamics and thermodynamics of binding[3,4], or to provide mechanistic insights such as revealing minor conformers[5,6].

Despite its utility, however, NMR presents a number of challenges to researchers. The assignment of each NMR signal to its corresponding atom in crowded NMR spectra can be an arduous process fraught with ambiguity. NMR experiments that characterize protein dynamics often require extensive acquisition time on expensive and unstable samples, and the analysis of dynamics data can be tedious and cumbersome.

In the following chapters, I present contributions to the field aimed at mitigating some of these challenges, as well as an application of such techniques to study an important class of enzymatic systems. I have improved covariance NMR techniques to assist in the assignment of proteins with crowded spectra. I have developed a software package to simplify and improve analysis of data from accordion relaxation spectroscopy experiments, which can be used to collect data more quickly than traditional experiments. And finally, I have applied NMR dynamics experiments to study the relationship between a non-ribosomal peptide synthetase carrier protein and its associated linker regions.

1.1 NMR assignment with 4D covariance correlation maps

Modified portions of this text have been published in the *Journal of Magnetic Resonance*[7] and submitted for publication in *Methods in Molecular Biology*[8].

A central task in NMR assignment involves relating nuclei to each other through either direct correlations or their mutual correlation to a common nucleus or set of nuclei. For example, protein amide H^N and N resonances are directly correlated to each other in 2D HN-HSQC spectra, forming (H^N, N) correlations. Similarly, methyl H^M and C^M resonances form (H^M, C^M) correlations in 2D HC-HSQC spectra. On the other hand, for large proteins, amide resonances are related to methyl resonances through their mutual correlation to C^α and C^β nuclei in 3D HNCA, HN(CA)CB and HMCMCBCA spectra[9,10].

Conventional assignment techniques approach assignment by first abstracting the raw data into a peak list made of frequency coordinates, a process known as peak-picking. Next, various frequency lists are compared to identify shared resonances. The results of this search are presented to users in the form of spectrum strips. Users are then tasked with accepting or rejecting proposed assignments based on comparisons between the strips. In the previous example, HNCA and HN(CA)CB spectra give rise to (H^N, N, C^α) and (H^N, N, C^β) correlations respectively, while HMCMCBCA spectra give rise to both (H^M, C^M, C^α) and (H^M, C^M, C^β) correlations for valine residues. Software algorithms can identify instances of common C^α and C^β frequencies among these coordinates and translate them into proposed assignments.

While effective, the conventional approach relies crucially on the premise that the frequency coordinates in each abstracted peak list are a faithful representation of the spectrum's true, underlying correlations. However, if this assumption is not fulfilled, the approach will fail outright. In such a case, only revising the peak list can rescue the assignment, yet this task is often fraught with ambiguity and error. In cases where signals have been erroneously picked or entirely neglected, the assignment suggestions offered by software will be incorrect or absent. In the former case, the mis-assignment of residues can have disastrous consequences including erroneous identification of structural constraints or binding sites. If signals have been neglected, the inability to extend a sequence of assigned residues can thwart entire projects. Furthermore, when searching for un-picked and often weak peaks, researchers may waste valuable time investigating potential ‘‘signals” that are in fact entirely noise. Because accurate peak-picking is absolutely essential, even small errors by experienced users can have disastrous consequences.

Over the past twelve years, covariance NMR has emerged as a complementary tool to conventional assignment techniques, because it does not rely on any form of abstraction. It accomplishes the same task as traditional approaches, namely relating unconnected nuclei through their mutual correlation to a common nucleus, but it does so by directly manipulating the raw data itself. The information provided separately by individual spectra are mathematically combined to create novel correlation maps that identify assignment candidates without making any *a priori* assumptions.

Although covariance NMR is a promising alternative to traditional methods, the technique has not been widely adopted by the NMR community. This is perhaps due to the high prevalence of false-positive signals in covariance spectra. In an effort to address this limitation, we have introduced pre- and post-processing steps to help reduce or eliminate such artifacts. We discuss these methods in chapter 3, and we introduce a particular application using 4D covariance spectra made possible with our procedures.

1.2 Measuring NMR relaxation rates with accordion spectroscopy

Modified portions of this text have been published in the *Journal of Biomolecular NMR*[11]

Measurement of NMR relaxation rates is a mainstay among the methods used to probe protein dynamics. Unfortunately, these measurements can require prohibitive acquisition times, on the order of a day or longer. Consequently, in the absence of *a priori* knowledge of the relevance of internal motions, the required investment in spectrometer time dampens the incentive to perform such experiments. Indeed, dynamics may prove to be important for a protein's function, but the return from lengthy measurements may be disappointing as well, for instance when only unstructured terminal regions display flexibility. To overcome costs in acquisition time, the accordion method[12] can be used to rapidly measure relaxation rates with accuracy indistinguishable from those measured with traditional methods[13–15]. In effect, the accordion method accelerates data

acquisition by encoding two indirect dimensions simultaneously. For example, when measuring relaxation rates, the traditional method involves recording a series of 2D HSQC experiments, each with a different relaxation period, effectively resulting in a 3D experiment. In the accordion version, the relaxation period is incremented synchronously with the period encoding signals with heteronuclear chemical shifts, reducing the dimensionality of the experiment from three to two. Whereas ten or more 2D experiments are typically recorded using the traditional method, the accordion method requires only two. Thus, relaxation rates can be estimated with a few hours of acquisition instead of days.

Despite its age, the NMR community has under-utilized accordion relaxation spectroscopy. Bodenhausen and Ernst first developed the accordion method for exchange spectroscopy in 1981[12], and Mandel and Palmer subsequently applied it to relaxation rates in 1994[13]. However, to our knowledge there have been only five publications[13–17] on accordion and relaxation analysis. This scarcity is even more surprising when one considers that the application of accordion spectroscopy to NMR relaxation is intuitively simple; the relaxation decay curve is encoded into the line-shape of NMR signals. Unfortunately, the various procedures available to extract the relaxation rates can be intimidating, time consuming and/or limited to resolved signals. The lack of discussion on the advantages and disadvantages of these procedures further impedes a popularization of the accordion technique. Thus, a unified software package enabling reliable and rapid data analysis is needed to promote the application and development of the accordion method.

In chapter 4 we present SARA (Software for Accordion Relaxation Analysis), a graphical, user-oriented software package designed to simplify and accelerate accordion data analysis. We have implemented existing protocols and designed novel ones allowing users to make use of the optimal fitting procedures for a given protein, bypassing the limitations of any one technique. In an effort to make it more accessible to occasional NMR users, SARA is written in MATLAB[18] and features an intuitive graphical user interface (GUI). In addition, we discuss the advantages and limitations of the analytical procedures implemented in SARA within the framework of their application to studies of protein dynamics. Finally, we describe a new protocol that employs all of the fitting procedures included in SARA and ensures critical inspection of the fitted data during accordion relaxation analysis. Overall, SARA provides an environment well-suited for routine analysis of relaxation rates obtained with accordion spectroscopy, a powerful technique capable of rapidly assessing protein dynamics.

1.3 Molecular cross-talk between an NRPS carrier protein and its linkers

It has been widely established that the prevalence of antibiotic resistant bacterial strains is increasing, most notably methicillin-resistant *Staphylococcus aureus* (MRSA)[19]. Vancomycin has been the treatment of choice for serious infections caused by MRSA; thus, reports of strains with reduced vancomycin susceptibility have generated great concern[20]. Vancomycin is largely synthesized by a non-ribosomal peptide synthetase (NRPS), and in fact a majority of the current and proposed treatment options for vancomycin resistant *Staphylococcus aureus*

strains are glycopeptides, lipopeptides, lipoglycopeptides, or streptogramins[20], all of which involve NRPS systems in their biosynthesis.

Non-ribosomal peptide synthetases are megadalton enzymatic assemblies found in bacteria and fungi that manufacture a number of important natural products: antibiotics, anticancer drugs, immunosuppressants, virulence factors, etc[21]. NRPSs use a remarkable, conserved, assembly line synthetic strategy[22]. This approach uses a modular architecture to add individual monomers to a growing NRP chain, where each NRPS module is responsible for the addition of a single substrate. A canonical module consists of at least three domains: a carrier protein, an adenylation domain and a condensation domain.

Carrier proteins (CP) are central to NRPS synthesis and serve to shuttle substrates between catalytic domains. First, phosphopantetheinyl transferases covalently attach a phosphopantetheine (PP) arm to a conserved serine residue of the CP. Adenylation (A) domains then load a substrate on to the PP arm in the form of a thioester linkage. Condensation (C) domains catalyze condensation between two substrates tethered to their respective carrier proteins. Alternatively, cyclization (Cy) domains replace C domains and catalyze both condensation and hetero-cyclization of serines and cysteines. The growing peptide chain is left attached to the C-terminal (downstream) carrier protein, and therefore synthesis proceeds from the N-terminus to the C-terminus of these supramolecular assemblies. The final module of an NRPS will typically contain a thioesterase (TE) domain, which catalyzes hydrolysis or macro-cyclization of the completed peptide.

Frequently, tailoring domains will also be present to catalyze modifications of the peptide, including methylation, oxidation, reduction and epimerization.

The modular nature of NRPS assembly lines lends itself well to the prospect of engineering novel “natural” products, including antibiotics. Ideally, antibiotic derivatives could be developed simply by modifying and swapping NRPS modules genetically[22]. While attempts at domain swapping based solely on bioinformatic techniques have been successful, synthesis occurs with efficiencies far lower than with cognate domains[21]. X-ray crystallographic studies of NRPSs have also failed to better motivate such efforts. Wild-type carrier proteins have been notoriously difficult to crystallize[23,24], and while structures of individual domains and pairs of domains have been solved[25], structures of multi-domain complexes have sometimes failed to reveal functional domain arrangements[23,26]. In fact, NMR structural studies have demonstrated that NRPSs do not form rigid structural complexes but rather operate through transient and sequential domain interactions[27,28].

Furthermore, it is well known that linker residues often do more than simply tether protein domains[29]. In NRPS and related PKS systems, it has been established that communication or COM domains, found at the termini of multi-domain enzymes, mediate interactions between modules on distinct polypeptide chains[30–32]. Investigators have also speculated on the role of linkers between domains in the same polypeptide chain[21]. Consequently, we sought to understand the role of linker residues in the dynamic interactions undergone by NRPS carrier proteins.

Specifically, we studied the influence of linker regions on the structure and dynamics of $PCP1_{ybt}$, the second carrier protein and first peptidyl carrier protein (PCP) from Yersiniabactin synthetase (YS). YS is an NRPS responsible for the synthesis of the virulence factor yersiniabactin (Ybt), a siderophore that scavenges iron from iron-depleted host environments. YS is found in: *Yersinia pestis*[33], the etiological agent of the plague; *Y.enterocolitica*[34], a food pathogen; and uropathogenic *E.coli*.[35], a major source of urinary tract infections. YS is a well-studied[33,36–43], hybrid non-ribosomal peptide synthetase – polyketide synthase (PKS) system composed of five modules arranged over four proteins.

In chapter 5, we show that communication exists between the core of $PCP1_{ybt}$ and its inter-domain linker regions. Using NMR we show that residues in the N-terminal linker interact with the domain core and modulate its dynamics. Further, by monitoring the protein's invisible unfolded state, we show that this linker region stabilizes the carrier protein fold.

2 Methods

2.1 Protein preparation

2.1.1 Cloning of $PCP1_{ybt}$

Fragments from the *Y. pestis* irp2 gene (Accession Number AAM85957) coding for residues 1383-1491, 1402-1482 and 1406-1482 of the protein HMWP2 (courtesy Dr. Christopher T. Walsh, Harvard Medical School) were PCR amplified using the primers PCP1_1383_KpnI_5p (5'-ACATATGGTACCATTGACTACCAGGC-3'), PCP1_1402_KpnI_5p (5'-CATATGGGTACCGCGGATTTACCCCAGGGC-3') and PCP1_1406_KpnI_5p (5'-CATATGGGTACCCAGGGCGACATTGAAAAACAGGTT-3') to introduce KpnI cut sites and PCP1_1482_EcoRI_3p (5'-GACGTCCCGGTCGAATGAGAATTCACCTGC-3') and PCP1_1491_EcoRI_3p (5'-ATATGTGAATTCTCAATCTTCAGGGG-3') to introduce an EcoRI cut site and stop codon.

The PCR products and target vector pET30a-GB1-TEV were digested with KpnI and EcoRI, gel purified and extracted. The fragments were ligated into pET30a-GB1-TEV to yield pET30a-GB1-TEV-PCP1-1383-1491, pET30a-GB1-TEV-PCP1-1402-1482 and pET30a-GB1-TEV-PCP1-1406-1482. The resulting plasmids direct production of residues 1383-1491, 1402-1482 and 1406-1482 of HMWP2 with an N-terminal GB1 tag followed by a hexahistidine tag and TEV cleavage site. The DNA sequence of each plasmid was confirmed by sequencing. Following TEV cleavage, a GT sequence remains at the N-terminus.

2.1.2 Expression of $PCP1_{ybt}$

Unless otherwise noted, the pH listed for each buffer is at 4 °C. All $PCP1_{ybt}$ constructs (pET30a-GB1-TEV-PCP1-1383-1491, pET30a-GB1-TEV-PCP1-1402-1482 and pET30a-GB1-TEV-PCP1-1406-1482) were expressed in identical manners.

pET30a-GB1-TEV-PCP1 vectors were transformed into competent E. coli BL21 (DE3) ΔEntD cells (courtesy Drs. Chalut and Guilhot, CNRS, Toulouse, France). A 5 ml culture of Luria broth (LB) with 50 µg/ml kanamycin was inoculated with a single transformed colony and grown at 37 °C and 250 rpm for 6 hours. 1 ml of the LB culture was added to 50 ml M9 minimal media (see section 2.1.5), with 1 g/L $^{15}NH_4$ Cl and/or 2 g/L ^{13}C glucose as the sole sources of nitrogen and/or carbon for labeled samples, and grown at 37 °C overnight. The 50 ml culture was added to the remaining 950 ml of media and growth continued at 37 °C. At an optical density of 0.6 at 600 nm (OD_{600}) , the cells were removed and cold-shocked on ice for 30 minutes. Next, isopropyl β-D-1 thiogalactopyranoside (IPTG) was added to 0.5 mM to induce protein expression, and growth continued overnight at 16 °C. Cells were harvested approximately 16 hours after induction at an optical density of approximately 1.1 by centrifugation at 4700 xg for 10 minutes.

2.1.3 Purification of $PCP1_{ybt}$

All $PCP1_{ybt}$ constructs (pET30a-GB1-TEV-PCP1-1383-1491, pET30a-GB1-TEV-PCP1-1402-1482 and pET30a-GB1-TEV-PCP1-1406-1482) were purified in identical manners.

The cell pellet was resuspended in 50 ml lysis buffer (50 mM Tris, pH 8.0, 500 mM NaCl, 25 mM imidazole, 0.1% Triton (w/v), 1 mg/ml lysozyme, 10 µg/ml DNase). Just before lysis, phenylmethanesulfonyl fluoride (PMSF) was added to a concentration of 2 mM. Lysis was performed with a Microfluidics 110Y microfluidizer, increasing the volume to approximately 100 ml. The lysate was clarified by centrifugation at 26,900 xg for 30 minutes and filtered with a 0.2 µm filter (Corning). The lysate was then loaded onto a 5 ml HisTrap HP column (GE Healthcare). The column was washed with 30 ml His. Buffer A (50 mM Tris, pH 8.0, 500 mM NaCl, 25 mM imidazole) and eluted with a linear gradient reaching 100% His. Buffer B (50 mM Tris, pH 8.0, 500 mM NaCl, 500 mM imidazole) over 20 column volumes (CV) at a flow rate of 2 ml/min using an Aktapurifier (GE Healthcare). Fractions containing GB1-TEV-PCP1 were identified by SDS-PAGE and pooled. One OD_{280} TEV protease was added per 20 OD_{280} sample to remove the GB1 and hexahistidine tags and the sample was dialyzed against 2L Dialysis Buffer (50 mM Tris, pH 8.0, 500 mM NaCl) overnight at 4 °C.

The dialyzed sample was again filtered with a 0.2 µm filter (Corning). Complete digestion of the sample to produce GB1-His6 and $PCP1_{ybt}$ was verified by SDS-PAGE. The dialyzed sample was loaded onto a 5 ml HisTrap HP column and washed with 30 ml Dialysis Buffer A. The flow-through containing $PCP1_{ybt}$ was

collected, concentrated using a 3K MWCO centrifugal filter (Millipore), and loaded onto a Superdex 75 16/60 pg column (GE Healthcare) that had been pre-equilibrated with 1.2 column volumes of NMR buffer (20 mM sodium phosphate, pH 6.60 at 22 °C, 150 mM NaCl, 1 mM DTT, 1 mM EDTA). Peak fractions were identified by SDS-PAGE and stored dilute at 4 °C.

Various concentrations of $PCP1_{ybt}$ were used in NMR samples, as indicated below. Concentration was determined by measuring absorbance at 280 nm (A_{280}) using an extinction coefficient of 8480 $M^{-1}cm^{-1}$ for $PCP1_{ybt}$ 1383-1491 or 6990 $M^{-1}cm^{-1}$ for $PCP1_{ybt}$ 1402-1482 and $PCP1_{ybt}$ 1406-1482. All NMR experiments were performed at 25 °C on a 600 MHz Bruker Avance III spectrometer with a QCI cryoprobe.

2.1.4 SEC-MALS of $PCP1_{ybt}$

To verify that the aberrant tumbling behavior observed in $PCP1_{ybt}$ 1383-1491 was the result of flexible linker residues rather than oligimerization, we measured the absolute molar mass of the protein in solution at high concentration using multi-angle light scattering (MALS). MALS data was collected after elution from a Superdex75 10/300 GL (GE Healthcare) analytical size exclusion chromatography (SEC) column. Two samples were run. The first was a 20 µL injection of a ^{15}N labeled, 760 µM sample. The second was a 200 µL injection of the same ^{15}N labeled, 760 µM sample.

2.1.5 Preparation of M9 minimal media

For 1 L of M9 minimal media, 6 g Na_2HPO_4, 3 g KH_2PO_4, and 0.5 g NaCl were combined and brought to 1 L with Milli-Q water. The pH was verified to be between 7.2 and 7.4. The media was then autoclaved for 30 minutes.

From sterile filtered stock solutions kept at 4 °C, 2 mL 1 M $MgSO_4$, 10 ml vitamin mix (see below), 2 ml solution Q (see below), 10 ml of 20% w/v glucose and 5 ml of 20% w/v ammonium chloride were added to the media. If ^{15}N or ^{13}C labeling was implemented, $^{15}NH_4Cl$ or ^{13}C-glucose was substituted as required. Finally, antibiotic(s) were added to appropriate concentrations.

For 1 L of vitamin mix, 0.5 g Thiamine-HCl, 0.1 g D-Biotin, 0.1 g Choline Chloride, 0.1 g Folic Acid, 0.1 g Niacinamide, 0.1 g D-Panthothenic Acid, 0.1 g Pyridoxal, and 10 mg Riboflavin were combined and brought to 1 L with Milli-Q water. Aliquots were stored at -20 °C.

For 1 L of solution Q, 5 g $FeCl_2$-$4H_2O$, 184 mg, $CaCl_2$-$2H_2O$, 64 mg H_eBO_3, 18 mg $CoCl_2$-$6H_2O$, 4 mg $CuCl_2$-$2H_2O$, 340 mg $ZnCl_2$, and 605 mg Na_2MoO_4-$2H_2O$ were combined and brought to 900 ml with Milli-Q water. Next, 8 ml of 5 N HCl was added before bringing the solution to 1 L with Milli-Q water. Aliquots were stored at at -20 °C.

2.1.6 Expression and purification of $Cy1_{ybt}$

The expression and purification of $Cy1_{ybt}$ in various labeling schemes has been described previously[44,45]. Two labeling schemes were used during data collection: one with uniform ^{2}H, ^{15}N and ^{13}C labeling (CDN) and the other with ^{1}H

and ^{13}C labeled methyl side-chains (δ1 position only) for residues Ile, Leu, and Val in an otherwise uniform ^{2}H, ^{15}N and ^{12}C background (ILV).

2.2 NMR experiments

2.2.1 $PCP1_{ybt}$ backbone assignment

Backbone assignment experiments for $PCP1_{ybt}$ 1383-1491 were conducted with a 1.4 mM ^{15}N and ^{13}C labeled sample. Six backbone experiments were performed: 3D HNCO, 3D HN(CA)CO, 3D HN(CO)CA, 3D HNCA, 3D HN(COCA)CB and 3D HN(CA)CB. All experiments were carried out using non-uniform sampling in the indirect dimensions. Schedules were generated using PoissonGap[46] at a sampling factor of 10%. The following table provides the number of scans for each FID in the spectrum as well as the spectral width (SW), carrier frequency (CAR), and number of uniform sampling points (N, complex points) or highest sampled point for non-uniform sampling (Max N, complex points) for each dimension. Spectra were reconstructed with the software istHMS[47] and processed with NMRPipe[48]. Assignment was performed manually with CARA[49]. Resonance assignments have been deposited in the BMRB (BMRB ID 30205).

		^{1}H			^{15}N			^{13}C		
Experiment	Scans	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	Max N	SW (PPM)	CAR (PPM)	Max N
HNCO	4	16	4.7	1024	31.25	117	64	11.5	176.7	75
HN(CA)CO	16	16	4.7	1024	31.25	117	64	11.5	176.7	75
HN(CO)CA	8	16	4.7	1024	31.25	117	64	30	55.2	100
HNCA	8	16	4.7	1024	31.25	117	64	30	55.2	100
HN(CO)CACB	16	16	4.7	1024	31.25	117	64	60	44.7	150
HN(CA)CB	16	16	4.7	1024	31.25	117	64	60	44.7	150

Table 2.1 Backbone acquisition parameters for $PCP1_{ybt}$ 1383-1491

No backbone assignment experiments were necessary for $PCP1_{ybt}$ 1402-1482 because of the minimal peak shifts in HN-HSQC spectra relative to $PCP1_{ybt}$ 1383-1491.

For $PCP1_{ybt}$ 1406-1482, the same set of backbone assignment experiments performed on $PCP1_{ybt}$ 1383-1491 were run on an 860 μM sample with ^{15}N and ^{13}C labeling. Identical NUS sampling schedules were used. On the other hand, the acquisition parameters were modified slightly to optimize the spectra for the sample. Spectra were reconstructed with the software NESTA-NMR[50].

		^{1}H			^{15}N			^{13}C		
Experiment	Scans	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	Max N	SW (PPM)	CAR (PPM)	Max N
HNCO	4	16	4.7	1024	28	117	64	9	177.2	75
HN(CA)CO	16	16	4.7	1024	28	117	64	9	177.2	75
HN(CO)CA	8	16	4.7	1024	28	117	64	28	55.7	100
HNCA	8	16	4.7	1024	28	117	64	28	55.7	100
HN(CO)CACB	16	16	4.7	1024	28	117	64	59	44.7	150
HN(CA)CB	16	16	4.7	1024	28	117	64	59	44.7	150

Table 2.2 Backbone acquisition parameters for $PCP1_{ybt}$ 1406-1482

2.2.2 $PCP1_{ybt}$ sidechain assignment

Three aliphatic sidechain assignment experiments were performed: 3D H(CCCO)NH, 3D (H)C(CCO)NH and 3D HC(C)H-TOCSY. All three experiments were acquired with uniform sampling. The following table contains the acquisition parameters for the spectra.

Experiment	Scans	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N
		^{1}H (detected)			^{15}N			^{1}H		
H(CCCO)NH	16	16	4.7	1024	31.25	117	20	6.4	4.7	50
		^{1}H (detected)			^{15}N			^{13}C		
(H)C(CCO)NH	16	16	4.7	1024	31.25	117	20	66	41.7	64
		^{1}H (detected)			^{13}C			^{1}H		
HC(C)H-TOCSY	8	16	4.7	1024	44	24.7	32	6.4	4.7	50

Table 2.3 Aliphatic sidechain acquisition parameters for PCP1$_{ybt}$ 1383-1491

Aromatic sidechain assignment was performed using 2D (HB)CB(CGCD)HD and 2D (HB)CB(CGCDCE)HE experiments. Both experiments were acquired with uniform sampling. The following table contains the acquisition parameters for the spectra.

		^{1}H			^{13}C		
Experiment	Scans	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N
(HB)CB(CGCD)HD	128	16	4.7	1024	22	32.7	28
(HB)CB(CGCDCE)HE	576	16	4.7	1024	22	32.7	28

Table 2.4 Aromatic sidechain acquisition parameters for PCP1$_{ybt}$ 1383-1491

2.2.3 PCP1$_{ybt}$ NOESY experiments

Three NOESY experiments were performed on PCP1$_{ybt}$ 1383-1491. The first experiment, a time-shared ^{15}N/^{13}C NOESY-HSQC, was performed using the same sample as used for the backbone and assignment experiments. This experiment was used during assignment, but was not used for structure calculations. The second experiment was a ^{15}N NOESY-HSQC performed using a 920 μM ^{15}N labeled sample. The final experiment was a ^{13}C HSQC-NOESY experiment performed using a 1.4 mM ^{15}N and ^{13}C labeled sample in D_2O. The

acquisition parameters are summarized in the table below. All experiments were performed using uniform sampling.

		^{1}H (INEPT)			^{15}N/^{13}C			^{1}H (NOESY)		
Experiment	Scans	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N
^{13}C-NOESY-HSQC (TS)	8	16	4.7	1024	44	24.7	64	11.4	4.7	45
^{15}N NOESY-HSQC (TS)	8	16	4.7	1024	31.25	117	64	11.4	4.7	45
^{15}N NOESY-HSQC	16	16	4.7	1024	28	117	49	11	4.7	190
^{13}C HSQC-NOESY	8	10	4.7	150	44.5	68.7	73	16	4.7	1024

Table 2.5 NOESY acquisition parameters for PCP1$_{ybt}$ 1383-1491

2.2.4 PCP1$_{ybt}$ ^{15}N relaxation experiments

^{15}N longitudinal and transverse relaxation rates, R_1 and R_2, were measured with a 500 μM sample of ^{15}N labeled sample. R_1 relaxation experiments were performed with relaxation times acquired in the following order: 0, 1071, 595, 238, 833, 476, 952, 357, 714 and 119 ms. R_2 relaxation experiments were performed with relaxation times acquired in the following order: 0, 333, 185, 74, 259, 148, 296, 111, 222 and 37 ms. R_2 relaxation experiments were performed with a CPMG pulse train to remove residual contributions to R_2 from any conformational exchange processes. The CPMG pulsing frequency was ν_{CPMG} = 758 Hz in all experiments.

The ^{15}N heteronuclear NOE experiment was performed with a 920 μM sample of ^{15}N labeled sample. It was acquired as a set of interleaved experiments with and without ^{1}H saturation during the recycling delay. The remaining acquisition parameters are summarized in the table below.

			^{1}H			^{15}N		
Experiment	Scans	Recycling delay (s)	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N
R_1	8	3	16	4.7	1024	31.25	117	128
R_2	8	3	16	4.7	1024	31.25	117	128
Het-NOE	64	5	16	4.7	1024	28	117	128

Table 2.6 ^{15}N relaxation acquisition parameters for $PCP1_{ybt}$ 1383-1491

The same set of experiments performed on $PCP1_{ybt}$ 1383-1491 were run on an 850 µM sample of ^{15}N labeled $PCP1_{ybt}$ 1402-1482 with identical acquisition parameters.

The same set of experiments performed on $PCP1_{ybt}$ 1383-1491 were run on a 600 µM sample of ^{15}N labeled $PCP1_{ybt}$ 1406-1482 with identical acquisition parameters.

2.2.5 NMR spectra of $Cy1_{ybt}$

NMR spectra of $Cy1_{ybt}$, a 52 kDa cyclization domain excised from the NRPS Yersiniabactin Synthetase[37], were recorded on at 25°C on an 800 MHz Varian spectrometer equipped with a Chili-Probe. The CDN and ILV samples (see section 2.1.6) were concentrated to 650 and 640 µM respectively in the final NMR buffer: 20 mM sodium phosphate pH 7.0, 10mM NaCl, 1mM EDTA, 5mM DTT and 5% D2O.

The NMR spectra were acquired with the following parameters. The HNCA and the HN(CO)CA were recorded on the ILV sample while the HNCO and the HN(CA)CO were recorded on the CDN sample. The HNCA was acquired with a 1s

recycling delay, 8 scans, and spectral parameters: ^{1}H (1200 complex points, 4.758 PPM carrier, 16000 Hz spectral width), ^{13}C (60 complex points, 58 PPM carrier, 5530 Hz spectral width), ^{15}N (48 complex points, 118 PPM carrier, 2836 Hz spectral width). The HN(CO)CA was acquired with a 1.1 s recycling delay , 32 scans, and spectral parameters: ^{1}H (810 complex points, 4.758 PPM carrier, 13500 Hz spectral width), ^{13}C (60 complex points, 58 PPM carrier, 6435 Hz spectral width), ^{15}N (33 complex points, 118 PPM carrier, 2836 Hz spectral width). The HNCO was acquired with a 1 s recycling delay, 8 scans, and spectral parameters: ^{1}H (825 complex points, 4.758 PPM carrier, 15001 spectral width), ^{13}C (53 complex points, 176.976 PPM carrier, 2816 Hz spectral width), ^{15}N (53 complex points, 118 PPM carrier, 2836 Hz spectral width). The HN(CA)CO was acquired with a 1 s recycling delay, 32 scans, and spectral parameters: ^{1}H (825 complex points, 4.758 PPM carrier, 15009 spectral width), ^{13}C (41 complex points, 176.976 PPM carrier, 2816 Hz spectral width), ^{15}N (52 complex points, 118 PPM carrier, 2836 Hz spectral width).

2.3 Data analysis

2.3.1 4D covariance script

Covariance processing with our script requires four principal components: a working installation of MATLAB or GNU Octave, the Covariance NMR Toolbox[51] by Snyder et al., the script itself and a set of 3D spectra. Links to download our script, the Covariance NMR Toolbox and GNU Octave are available on our website at http://frueh.med.jhmi.edu/software-downloads/.

Our script has been tested in both MATLAB 8.3+ (2014+)[18] and GNU Octave 4.0+[52] on Linux, Mac and Windows operating systems. However, due to the simplicity of the code, it is anticipated that our script will run successfully on most modern versions of either MATLAB or Octave. Users are directed to their institution for MATLAB licensing. Octave is available free of charge. The Covariance NMR Toolbox should be decompressed, placed in a permanent location on the disk and added to the MATLAB or Octave path with the **addpath** command. The script itself should be copied and edited for each separate instance of its execution, much like NMRPipe processing scripts. The input 3D spectra should be in NMRPipe format, but they may be in either plane-by-plane format or in a single, monolithic file.

2.3.2 $PCP1_{ybt}$ structure calculations

Assignment of NOESY cross-peaks was performed manually using CARA. We assigned 1,716 unambiguous distance constraints. In addition, 146 angle constraints were obtained with TALOS-N[53]. Structure calculations were performed using CYANA[54,55] version 2.1. For the final structure calculation, 100 structures were calculated using 50,000 steps. The 20 structures with the lowest target function were chosen for water refinement in explicit solvent using CNS[56]. The NMR ensembles were analyzed with the protein structure validation suite PSVS[57], which includes PROCHECK_NMR[58] and MolProbity[59]. The structure bundle has been deposited in the PDB (PDB ID 5U3H).

2.3.3 ^{15}N CEST experiments

^{15}N CEST experiments were performed identically on both PCP1$_{ybt}$ 1383-1491 and PCP1$_{ybt}$ 1406-1482. ^{15}N CEST experiments on PCP1$_{ybt}$ 1383-1491 were run with a 1.4 mM ^{15}N labeled sample while those on PCP1$_{ybt}$ 1406-1482 were run with a 600 μM ^{15}N labeled sample. The pulse sequence was implemented as described previously[60]. For each construct, two sets of CEST experiments were performed at nominal ^{15}N B_1 fields of 25 Hz and 12.5 Hz. The B_1 field was scanned across 64 points of the ^{15}N spectral width, from 131 to 103.438 PPM. The remaining acquisition parameters are given in the table below.

			^{1}H			^{15}N		
Experiment	Scans	Recycling delay (s)	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N
^{15}N CEST	8	1	16	4.7	1024	28	117	64

Table 2.7 ^{15}N CEST acquisition parameters

Calibration of the ^{15}N B_1 fields was performed as described previously[60]. The ^{15}N carrier was centered on a peak that was well separated in the ^{1}H dimension. A 2D experiment was performed where the length of the ^{15}N B_1 field was varied during t_1. For the 25 Hz field, 100 real points were acquired in t_1 with a sampling frequency of 100 Hz. For the 12.5 Hz field, 50 real points were acquired in t_1 with a sampling frequency of 50 Hz. A single t_1 slice centered on the peak of interest was extracted from the interferogram and Fourier transformed. The B_1 field profile was fit with a Gaussian distribution, and its mean and standard deviation were used in ChemEx during CEST analysis (see below).

The data was processed with NMRPipe. Peak heights were extracted by fitting the series of spectra with the NMRPipe program nlinLS. CEST profiles were fit with the software ChemEx[60] to extract the exchange parameters.

2.3.4 ^{15}N CPMG relaxation dispersion experiments

^{15}N CPMG relaxation dispersion (RD-CPMG) experiments were run on a 1.4 mM ^{15}N labeled sample of PCP1$_{ybt}$ 1383-1491 at a single field of 600 MHz. The pulse sequence was implemented as described previously[61]. A reference experiment was acquired with a relaxation delay of 0 ms. The remaining experiments were acquired with a relaxation delay of 40 ms and ν_{CPMG} frequencies: 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 350, 400, 500, 600, 700, 750, 800, 900, and 1000 Hz. Selected repeat experiments were acquired for error analysis, including the reference experiment and those with ν_{CPMG} frequencies of 25, 150, 500, 750, and 1000 Hz (see below). The remaining acquisition parameters are summarized in the table below.

			^{1}H			^{15}N		
Experiment	Scans	Recycling delay (s)	SW (PPM)	CAR (PPM)	N	SW (PPM)	CAR (PPM)	N
RD-CPMG	8	4	16	4.7	1024	29.5	117	128

Table 2.8 ^{15}N RD-CPMG acquisition parameters

Peak heights in each spectrum were fit relative to their corresponding height in the reference spectrum using the NMRPipe program nlinLS. This approach directly fit the peak height ratio I/I_0 for each ν_{CPMG} frequency, where I is the peak height in the CPMG pulsing experiment and I_0 is the peak height in the reference experiment. For each residue, the largest observed deviation in I/I_0 across all of the repeat experiments was taken to be the error σ_{I/I_0}. This value was then applied

to all ν_{CPMG} frequencies for the residue. Peak height ratios were converted to effective R_2 relaxation rates $R_{2,eff}$ using the equation

$$R_{2,eff} = \frac{-\ln\left(\frac{I}{I_0}\right)}{T} \tag{2.1}$$

where T represents the relaxation delay (40 ms). Rates were then corrected for evolution outside of the transverse plane with the equation

$$R_{2,eff}^{corrected} = \frac{\tau_{CP} \cdot R_{2,eff} - \frac{1}{2} R_1 \tau_\pi \sin^2(\Phi)}{\tau_{CP} - \frac{1}{2} R_1 \tau_\pi \sin^2(\Phi)} \tag{2.2}$$

where $\tau_{CP} = \frac{1}{2\nu_{CPMG}}$, $\Phi = \frac{\tau_{CP}}{2} \cdot \Delta\omega_{CAR}$, $\Delta\omega_{CAR}$ is the offset of the residue from the ^{15}N carrier, τ_π is the ^{15}N 180° pulse width (83 µs) and $R_1 = 1.4\ s^{-1}$ [62]. Peak height ratio errors were propagated to relaxation rate errors using the equation

$$\sigma_{R_{2,eff}} = \frac{\sigma_{I/I_0}}{T \cdot \left(\frac{I}{I_0}\right)} \tag{2.3}$$

We fit the relaxation dispersion profiles with the program relax[63]. Relax fits the profiles to different mathematical models, each making different assumptions about the underlying exchange process. It then chooses the best model for each residue based on the corrected Akaike information criterion.

2.3.5 Model Free analysis

Model Free analysis of the R_1, R_2 and heteronuclear NOE data was performed with the program ROTDIF[64,65]. ROTDIF fits both the global rotational diffusion tensor as well as the individual values of S^2 and other model-free parameters at each residue. For all constructs, the H^N-N bond orientations used

during analysis were taken from the first structure of the water refined $PCP1_{ybt}$ 1383-1491 bundle. It is recommended that residues having a value of S^2 less than 0.75 or a value of R_{ex} greater than 10% of R_2 be removed from the fit of the global diffusion tensor. This was accomplished by iteratively performing the Model Free analysis and removing residues not satisfying the requirements. Predicted values for the rotational diffusion time τ_c based on molecular weight were calculated with the software COAST[66].

In $PCP1_{ybt}$ 1383-1491, the following residues were removed when fitting the global rotational diffusion tensor: 1383-1403, 1417-1418, 1421, 1423-1426, 1433, 1439, 1446-1447, 1456-1458, 1461, 1474, and 1476-1491. An axially symmetric diffusion tensor best fit the data, with τ_c = 8.54 ns, yet the predicted value based on molecular weight is 6.6 ns. This discrepancy is discussed in section 5.2.3.

In $PCP1_{ybt}$ 1402-1482 the following residues were removed when fitting the global rotational diffusion tensor: 1403-1404, 1417-1418, 1421, 1423-1426, 1428, 1432-1433, 1438, 1446, 1455-1456, 1458, and 1476-1482. An axially symmetric diffusion tensor best fits the data, with τ_c = 5.17 ns. The predicted value for τ_c based on molecular weight is 5.1 ns.

In $PCP1_{ybt}$ 1406-1482, the following residues were removed when fitting the global rotational diffusion tensor: 1406-1408, 1416, 1418-1419, 1421, 1423-1426, 1428-1433, 1435-1436, 1443, 1452, 1456, 1458, 1463-1464, 1467-1468, 1472-1473, and 1478-1482. An axially symmetric diffusion tensor best fits the data, with τ_c = 4.78 ns. The predicted value for τ_c based on molecular weight is 4.8 ns.

3 NMR assignment with 4D covariance correlation maps

Modified portions of this text have been published in the *Journal of Magnetic Resonance*[7] and submitted for publication in *Methods in Molecular Biology*[8]

3.1 Covariance NMR Theory

Covariance NMR was developed in 2004 by Brüschweiler and coworkers[67–71] as an alternative to conventional Fourier transform processing. Initially, spectra were covaried with themselves as a means to enhance resolution in indirectly detected dimensions. Soon after, Blinov et al. introduced unsymmetrical covariance NMR[72,73], whereby novel correlation maps could be created by covarying two *different* spectra. Later, related techniques broadened to include hyperdimensional NMR[74,75], assignment without peak lists[76] and cross-spectra[77]. Finally, Snyder at al. codified the theory of covariance NMR[78,79] and expanded its applications to higher dimensionality spectra[80,81].

3.1.1 Principles of Covariance NMR

Unsymmetric covariance NMR links distinct, unconnected nuclei to each other through their mutual correlation to a common nucleus. Unsymmetric covariance spectra are formed from the matrix product of two different NMR spectra. For example, a spectrum **A** of (I, K) correlations can be covaried with a spectrum **B** of (L, K) correlations to form a covariance spectrum **C** of (I, L) correlations, calculated as

$$C(i,l) = \sum_{k=1}^{K} A(i,k) \cdot B(l,k) \tag{3.1}$$

or using matrix notation

$$\boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{B^T} \tag{3.2}$$

Figure 3.1 illustrates such a matrix product. If the two input spectra share a common signal along the subsumed (K) dimension, such as in panel (b), then the resulting covariance spectrum will feature a correlation at the corresponding row and column of the signals in the first and second spectrum respectively. However, if the two spectra do not share a common signal in the subsumed dimension, as in panel (c), then the covariance spectrum will be empty. False-positive artifacts originate from covariance between 1D slices that share a signal purely by coincidence. Such an example is shown in panel (d). Here, each original spectrum contains two peaks, marked with * and †, that are nearly degenerate in the K dimension. As expected, the covariance spectrum features the desired correlations at the ** and †† positions. However, as a result of near degeneracy, it also features undesirable peaks at the *† and †* positions. Such artifacts are an inevitable result of traditional covariance methods and are especially prevalent when dealing with crowded spectra, for example those of large, intrinsically disordered and some α-helical proteins. Panel (d) also underscores the importance of carefully processing the subsumed dimension. Any signals shared between the input spectra, even those originating from spurious signals or artifacts, will be carried over to the covariance spectrum.

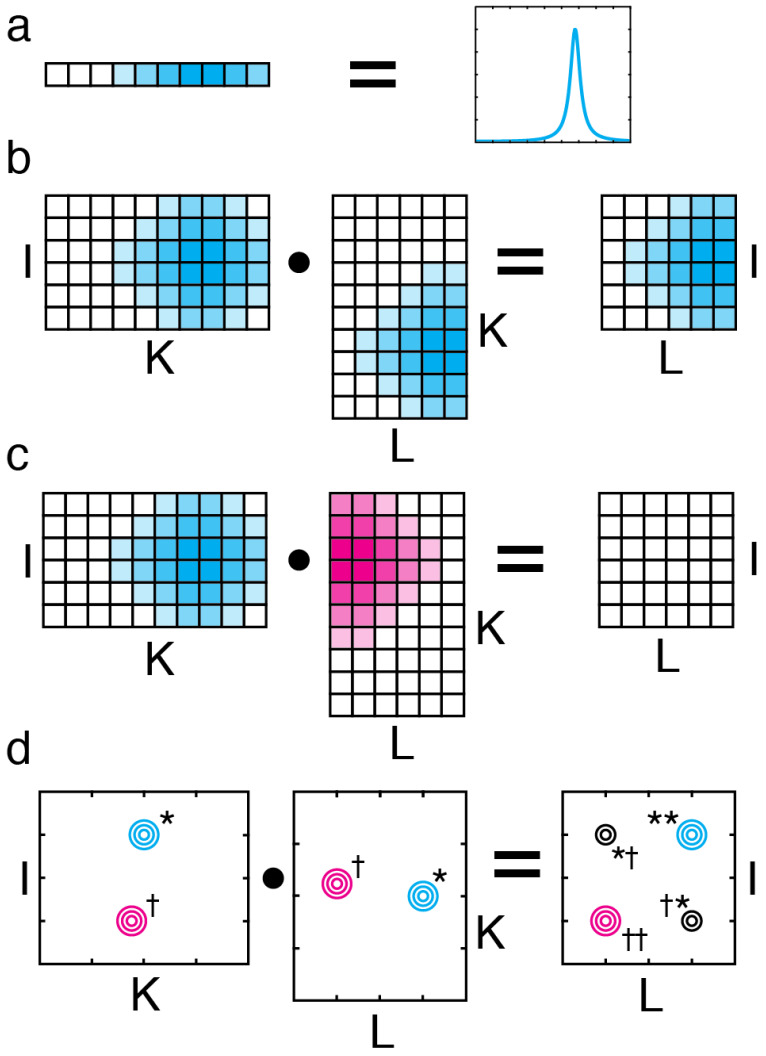

Figure 3.1 Unsymmetrical covariance is equivalent to matrix multiplication (a) The series of colored squares to the left represent the peak plotted on the right as a row vector. The intensity of the color in each square corresponds to the amplitude of the peak. (b) 2D spectra can be represented as matrices. An I x K matrix multiplied by a K x L matrix results in an I x L matrix, subsuming the K dimension. If two spectra share a signal along the K dimension, there will be a peak in their resulting matrix product. (c) If two spectra do not share a peak in the K dimension, their matrix product will be devoid of any peaks. (d) If two sets of peaks are nearly degenerate in the K dimension, the covariance spectrum will feature false-positive artifacts. Here, each pair of peaks in the input spectra, marked with † or *, produces a peak in the covariance spectrum. The peaks marked †† and ** are the desired correlations, while the peaks marked †* and *† are false-positive artifacts.

3.1.2 Covariance with spectral derivatives

Our lab has introduced a pre-processing step that helps to reduce the number and severity of false-positive artifacts in covariance spectra. The matrix product operation at the heart of covariance NMR can be broken down into a series of inner products between slices from the spectra. Figure 3.2 illustrates that the inner product is further separable into two operations, element-wise multiplication and summation. With standard covariance NMR processing (panels (a-f)), if two 1D slices feature even partially overlapping signals, then their element-wise product will be strictly positive, and the subsequent summation will inevitably produce an incorrect and undesirable peak in the covariance spectrum. However, if we first take the derivative of each slice (panels (g, h)), prior to element-wise multiplication, the product will be strictly positive only in the case of exact alignment between the signals (panels (i-k)). If the signals are not perfectly aligned (panels (l-n)), the element-wise product will feature both positive and negative signals, and after summation, the negative signals will act to reduce, change the sign, or eliminate false-positive correlations. Spurious negative correlations can then be eliminated by restricting the spectrum to only its positive elements.

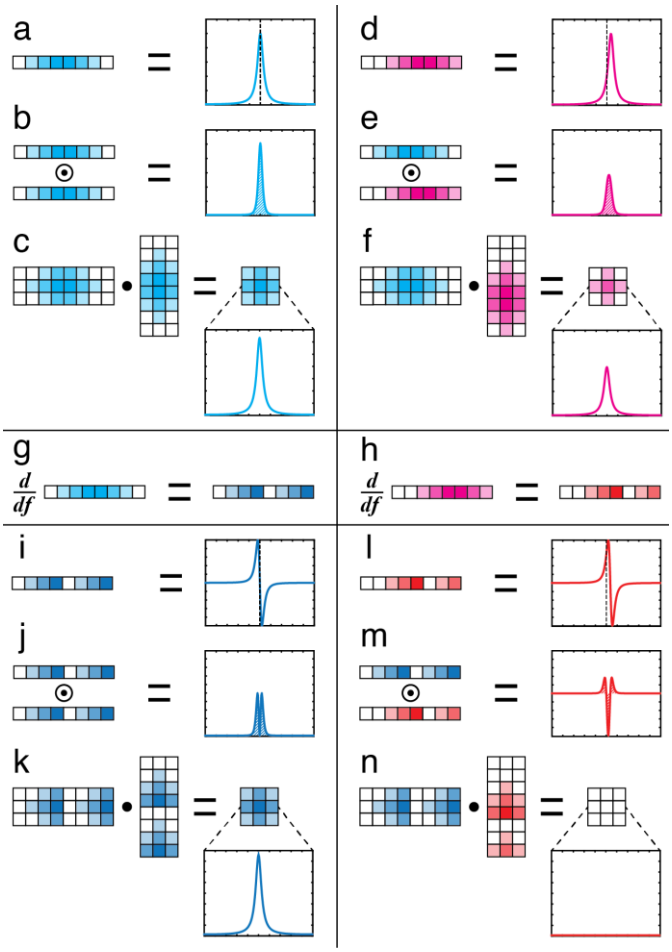

Figure 3.2 Derivatives and covariance spectra. (a) The series of colored squares to the left represents the peak to the right as a row vector. The intensity of the color in each square corresponds to the amplitude of the peak. The peak is centered on the dotted line. (b) When two peaks align perfectly, the element-wise product of their row vectors is strongly positive. (c) The matrix product of two 2D spectra is a covariance spectrum. Each of its elements is a sum of an element-wise product, like that shown in (b). In this case, a strong signal is produced in the covariance spectrum. (d) A different peak, slightly offset from the peak in (a), is represented by a shifted row vector. (e) The shift between the peaks in (a) and (d) reduces their element-wise product. (f) The two signals from (a) and (d) still produce an undesirable peak in the covariance spectrum. (g, h) Taking the derivatives of the vectors in (a) and (d) yields the vectors in (i) and (l) respectively. The colored squares have been reordered for clarity, even though they no longer correctly represent the changes in intensity. (i) The derivative of the signal in (a) is still centered on the dotted line. (j) If two peaks align perfectly, then their derivatives will as well, and the element-wise product of the two vectors is strictly positive. (k) Upon summation, the element-wise product produces a strong peak in the covariance spectrum, as desired. (l) The derivative of the signal in (d) is offset from the dotted line. (m) The element-wise product of the vectors from (i) and (l) has both positive and negative values. (n) Upon summation, the positive and negative values from the element-wise product cancel each other, producing an empty covariance spectrum. The false-positive artifact has been eliminated.

Given an analytical expression for the line shape of signals in the subsumed dimension, we can derive an equation for the amplitude of false positive artifacts as a function of the frequency separation between the peaks. For example, assuming a Lorentzian line shape, the amplitude of the covariance peak between two signals can be calculated as follows. Let the function L represent a Lorentzian line shape parameterized by an amplitude A, a relaxation rate R, and a frequency offset f_0:

$$L(A, R, f_0) = \frac{A \cdot R}{(R^2 + 4\pi^2(f - f_0)^2)} \tag{3.3}$$

Taking the derivative of L as a function of frequency f gives

$$\frac{dL(A, R, f_0)}{df} = \frac{-8\pi^2 f \cdot A \cdot R}{(R^2 + 4\pi^2(f - f_0)^2)^2} \tag{3.4}$$

Letting the labels A and B represent peaks from input spectra **A** and **B** respectively, the continuous frequency analog of equation 3.1 with Lorentzian line shapes is

$$C = \int_{-\infty}^{\infty} \frac{dL(A_A, R_A, f_A)}{df} \cdot \frac{dL(A_B, R_B, f_B)}{df} df \tag{3.5}$$

where C denotes the amplitude of a correlation in a covariance spectrum at a particular position and corresponds to C(i, l) in equation 3.1. Without loss of generality, we can set the frequency of the peak from **A** equal to zero and parameterize the peak from **B** in terms of its offset relative to that of **A**, Δf. With this simplification equation 3.5 reduces to

$$C = \int_{-\infty}^{\infty} \frac{dL(A_A, R_B, 0)}{df} \cdot \frac{dL(A_B, R_B, \Delta f)}{df} df \tag{3.6}$$

From this equation, the amplitude of the covariance peak is calculated as

$$C = \frac{4\pi^2 A_A A_B (R_A + R_B)((R_A + R_B)^2 - 12\pi^2 \Delta f^2)}{((R_A + R_B)^2 + 4\pi^2 \Delta f^2)^3} \tag{3.7}$$

Figure 3.3 (a) plots C as a function of Δf. The signal is maximum when both peaks are found at the same frequency (i.e. Δf = 0), and the amplitude is

$$C = \frac{4\pi^2 A_A A_B}{(R_A + R_B)^3} \tag{3.8}$$

As Δf grows, the covariance peak shrinks until it becomes zero at

$$\Delta f_{crit} = \frac{1}{\sqrt{3}} \frac{(R_A + R_B)}{2\pi} = \frac{1}{\sqrt{3}} FWHM_{av} \tag{3.9}$$

where $FWHM_{av}$ is the average of the two peaks' full width at half max. Frequency differences less than Δf_{crit} result in positive covariance peaks, whereas differences greater than Δf_{crit} produce negative artifacts in the covariance spectrum. Once the signals are completely resolved, no spurious correlation is observed. Because negative artifacts can easily be identified and ignored by retaining only positive elements after calculation, the derivative procedure effectively eliminates all artifacts that originate from degeneracies greater than or equal to Δf_{crit}.

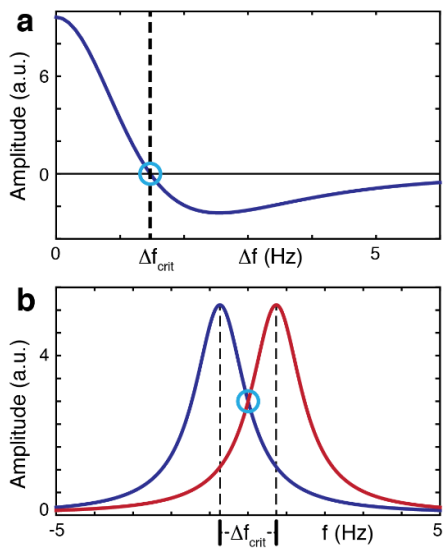

Figure 3.3 False positive artifacts in covariance spectra (a) Covariance peak amplitude C as a function of frequency offset Δf. Here $A_A = A_B = 1$ and $R_A = R_B = 8\ s^{-1}$. Δf_{crit} marks the zero crossing of the covariance peak (b) An illustration of two Lorentzian line shapes with their inflection points in alignment. For Lorentzians, this condition is equivalent to a frequency separation of Δf_{crit}

Incidentally, Δf_{crit} is also the frequency offset at which the left inflection point of one Lorentzian line shape aligns with the right inflection point of the other. Because this condition is easily identified by visual inspection (Figure 3.3 (b)), we suggest this criterion as a rule of thumb for anticipating when two signals will no longer produce a positive peak in the covariance spectrum. Lorentzian line-shapes correspond to Fourier transforms of free induction decays (FIDs) that have fully relaxed. However, this is most often not the case in the indirect dimensions of 3D spectra, and FIDs must be apodized before Fourier transformation. For line shapes less heavy-tailed than the Lorentzian distribution, we anticipate that artifact elimination will occur at frequency differences less than those that are required to align the inflection points. By repeating the derivation described for Lorentzian line

shapes with Gaussian line-shapes parameterized by amplitude A, line width parameter σ, and frequency offset f_0,

$$G(A, \sigma, f_0) = Ae^{-\frac{4\pi^2(f-f_0)^2}{2\sigma^2}} \tag{3.10}$$

one obtains

$$\Delta f_{crit} = \frac{1}{2\pi}\sqrt{\sigma_A^2 + \sigma_B^2} \tag{3.11}$$

The inflection points of Gaussian line shapes align at

$$\Delta f = \frac{\sigma_A + \sigma_B}{2\pi} \tag{3.12}$$

Thus, because $\sqrt{\sigma_A^2 + \sigma_B^2}$ is strictly less than $\sigma_A + \sigma_B$ for positive σ, the zero crossing of partial degeneracy artifacts must occur prior to the alignment of the inflection points.

3.1.3 Element-wise multiplication of covariance spectra

In addition to the improvements offered by the derivative pre-processing step, we have also implemented post-processing enhancements to the standard covariance protocol. Because we are often able to find more than one common nucleus through which we can relate unconnected nuclei, we can combine covariance spectra obtained through different nuclei to further decrease false-positive artifacts. Figure 3.4 demonstrates that two nuclei I and L can be related to each other through common correlations to two different nuclei K_1 and K_2. A spectrum of (I, K_1) correlations can be covaried with a spectrum of (L, K_1) correlations to form a covariance spectrum of (I, L) correlations, and spectra with

(I', K_2) and (L', K_2) correlations create a covariance spectrum of (I', L') correlations. Here, K_2 represents a completely different nucleus than K_1, e.g. C^{β} rather than C^{α}, whereas the apostrophes indicate a second instance of the same nucleus, e.g. I and I' are both H^N. Although both the (I, L) and (I', L') spectra may each contain false-positive signals originating from truly degenerate K_1 or K_2 frequencies, the artifacts in each spectrum will likely be different. As a result, taking the element-wise product of the two covariance maps will reinforce the shared peaks while reducing erroneous signals. Furthermore, if we restrict ourselves to only the positive correlations from each individual covariance map prior to element-wise multiplication (see above), we eliminate any possibility of introducing erroneous signals through the multiplication of two negative peaks.

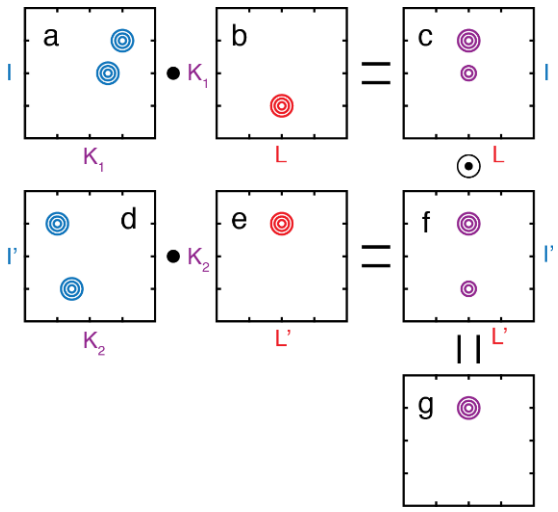

Figure 3.4 The element-wise product eliminates artifacts Spectrum (a) is covaried with spectrum (b) to produce spectrum (c). The dot symbol indicates matrix multiplication/covariance. The top-right signal in spectrum (a) gives rise to the top signal in spectrum (c). The second signal in spectrum (a) does not align perfectly with the signal in spectrum (b) in the subsumed dimension, and as a result, gives rise to the smaller, artifact signal in spectrum (c). Similarly, spectra (d) and (e) are covaried to produce spectrum (f). The small, artifact peak in spectrum (f) is again the result of covariance between imperfectly aligned signals in (d) and (e). (g) The element-wise product of (c) and (f) reinforces the correct peak while eliminating the artifacts. We represent the element-wise product operator with a circled dot.

3.1.4 Four-dimensional covariance spectra

Our applications so far have primarily focused on covarying two 3D spectra with correlations of the form (I, J, K) and (L, M, K) to create a 4D spectrum with (I, J, L, M) correlations (Figure 3.5). A 4D covariance spectrum of this type amounts to the inner product of every possible combination of 1D slices along the K dimension of each spectrum. As a result, all of the previously discussed improvements remain applicable. The derivative can be used as a pre-processing step before covariance, and the element-wise product of multiple 4D spectra can be used after covariance to combine spectra and reduce or eliminate artifacts.

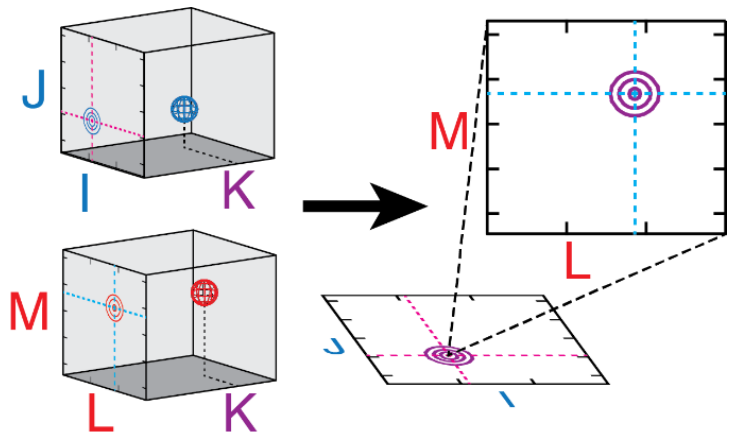

Figure 3.5 4D covariance spectra A 3D spectrum of (I, J, K) correlations can be covaried with a second 3D of (L, M, K) correlations to form a 4D covariance spectrum with (I, J, L, M) correlations. If the two 3D spectra share a signal at a common frequency in the K dimension, as shown, there will be a peak in the covariance spectrum at the corresponding frequencies in the I, J, L and M dimensions. A 4D spectrum can be visualized as a plane of planes. As a result, each point in the I/J plane corresponds to an entire L/M plane.

Typically these 4D spectra map one set of "spin anchors" to another set of "spin anchors" through an intermediary set of nuclei. Here we borrow the term spin anchor from the software application CARA[49] to refer to a correlation between a hydrogen nucleus and its directly attached, heavy atom nucleus. Much of the NMR

assignment process revolves around spin anchors. They are the correlations observed in HN- and HC-HSQC spectra, and many multidimensional NMR experiments correlate a spin anchor to some other nucleus or nuclei. For example, an HNCA correlates the H^NN anchor with C^α and an HMCMCBCA correlates valine methyl H^MC^M anchors with C^α and C^β. 4D spectra correlating spin anchors are particularly useful and simple to conceptualize. They are easy to navigate, and they lend themselves well to assignment without peak lists.

3.2 Using our 4D covariance processing script

3.2.1 Preparing spectra

Before using our script to calculate a 4D covariance spectrum, users must first prepare their 3D spectra appropriately. To successfully use our script: each spectrum must be transposed correctly; the digital resolution must match appropriately among all 3D spectra; and the spectra must have a self-consistent calibration.

Our script requires that the spin anchor nuclei be placed in the X and Y dimensions of the NMRPipe spectrum and that the shared nucleus be placed in the Z dimension. The Z dimension corresponds to the K dimension in Figure 3.5 and is the dimension that will be subsumed during covariance. To satisfy this requirement, users must transpose their spectra appropriately in NMRPipe. This can be accomplished using any combination of the **nmrPipe** functions TP and ZTP as well as when reading (**xyz2pipe**) and writing (**pipe2xyz**) data with the options -x, -y and -z. Table 3.1 describes the effect of each function or option when reading

the data from the disk, transforming it in memory and writing back to the disk. Many combinations amount to the same result. For example, to transfer data from the Y dimension to the Z dimension, use a combination of **xyz2pipe** -y and **pipe2xyz** -z. The same could be accomplished with **xyz2pipe** -z and **pipe2xyz** -x or even **xyz2pipe** -x, **nmrPipe** -fn TP, **nmrPipe** -fn ZTP and **pipe2xyz** -x. Users can verify that the data is correctly transposed by using the command **showhdr**. The shared nucleus should be listed under the column "Z-Axis" for each spectrum.

Function/option	Before	After	Start -> End
xyz2pipe -x	ABC	ABC	Disk -> Memory
xyz2pipe -y	ABC	BAC	Disk -> Memory
xyz2pipe -z	ABC	CAB	Disk -> Memory
nmrPipe -fn TP	ABC	BAC	Memory -> Memory
nmrPipe -fn ZTP	ABC	CBA	Memory -> Memory
pipe2xyz -x	ABC	ABC	Memory -> Disk
pipe2xyz -y	ABC	BAC	Memory -> Disk
pipe2xyz -z	ABC	BCA	Memory -> Disk

Table 3.1 Data transposition with NMRPipe. A, B and C each represent a dimension name. The order of the three characters in each box indicates the order of the data in the dimensions X, Y and Z of the spectrum respectively. The right-most column specifies the location of the data before and after using the function/option.

If multiple pairs of 3D spectra will be covaried and later multiplied together, then the order of the spin anchor dimensions should be consistent across each pair of 3D spectra. For example, if H^NN spin anchors in HNCA and HN(CA)CB spectra will be connected to H^MC^M spin anchors in HMCMCBCA spectra through both C^α and C^β, then H^N should be found in the same dimension of both the HNCA and HN(CA)CB spectra, e.g. along the X dimension. The same is true for the other three nuclei comprising the two spin anchors that will be correlated. Note that once the shared nucleus has been properly placed in the Z dimension, correctly placing

the remaining nuclei should require *at most* an XY transpose with the **nmrPipe** function TP or xyz2pipe and pipe2xyz options -y.

The next step when preparing spectra is to match the digital resolution along any dimensions that undergo covariance or element-wise multiplication. This is essential to ensure that each discrete element of one spectrum is multiplied with its corresponding element in the other spectrum. As discussed above, element-wise multiplication is part of both the covariance operation and our post-processing scheme. If a 4D spectrum is calculated from a pair of 3D spectra with correlations of the form (I, J, K_1) and (L, M, K_1), as in Figure 3.5, then the digital resolutions of the two K_1 dimensions must be matched before covariance. Furthermore, if there is a second pair of 3D spectra with correlations (I', J', K_2) and (L', M', K_2), then not only must the digital resolutions of the two K_2 dimensions match, but the resolution of I must match that of I', J must match J', L must match L', and M must match M'. Stated differently, the digital resolutions of each K dimension must match *within each pair* of covaried 3D spectra, and the digital resolutions of the I, J, L and M dimensions must match respectively *across all pairs* of 3D spectra.

Figure 3.6 illustrates why and how the digital resolutions are matched. Specifically, we consider spectra with different spectral widths, carrier frequencies, and initial resolutions. Panel (a) represents a slice of 120 points along a ^{13}C dimension centered at 55 PPM with a spectral width of 30 PPM. Panel (b) represents a slice of 96 points centered at 58 PPM with a spectral width of 28 PPM. Both spectra contain a signal at 52 PPM. Beneath each spectrum is a series of dots representing the points at which each spectrum is sampled. Each space

between dots represents 8 points of the spectrum. Panel (c) demonstrates that although the two spectra represent the same underlying signal, they sample that signal at very different data coordinates. Panel (d) shows that if we compress the second spectrum to align each of its points with the first spectrum, the signals themselves no longer align. As the two spectra are currently processed, there is simply no way to correctly multiply them in an element-wise fashion. The solution is simply to zero pad the two spectra to reach the same digital resolution after Fourier transformation. In this case, the first spectrum contains 120 points over a spectral width of 30 PPM, giving a digital resolution of 4.00 points per PPM, whereas the second spectrum features 3.43 points per PPM. Zero padding the second spectrum from 96 points to 112 points, will increase its digital resolution to 4.00 points per PPM, matching the first. It is important that the digital resolutions match exactly; even small differences can still give rise to errors. For example, a digital resolution of 4.00 points per PPM and a spectral width of 30 PPM would result in a spectrum of 120 points. However, a digital resolution of 4.02 points per PPM and the same spectral width of 25 PPM would result in a spectrum of approximately 120.6 points, which would be rounded up to 121 points. The newly zero padded spectra in panels (e) and (f) are overlaid in panel (g). Now both spectra sample the underlying signal at the same locations. Finally, the limits of each spectrum are restricted to a common region, here 44 to 70 PPM as in panel (h), and element-wise multiplication is well defined. This last step is performed within our processing script.

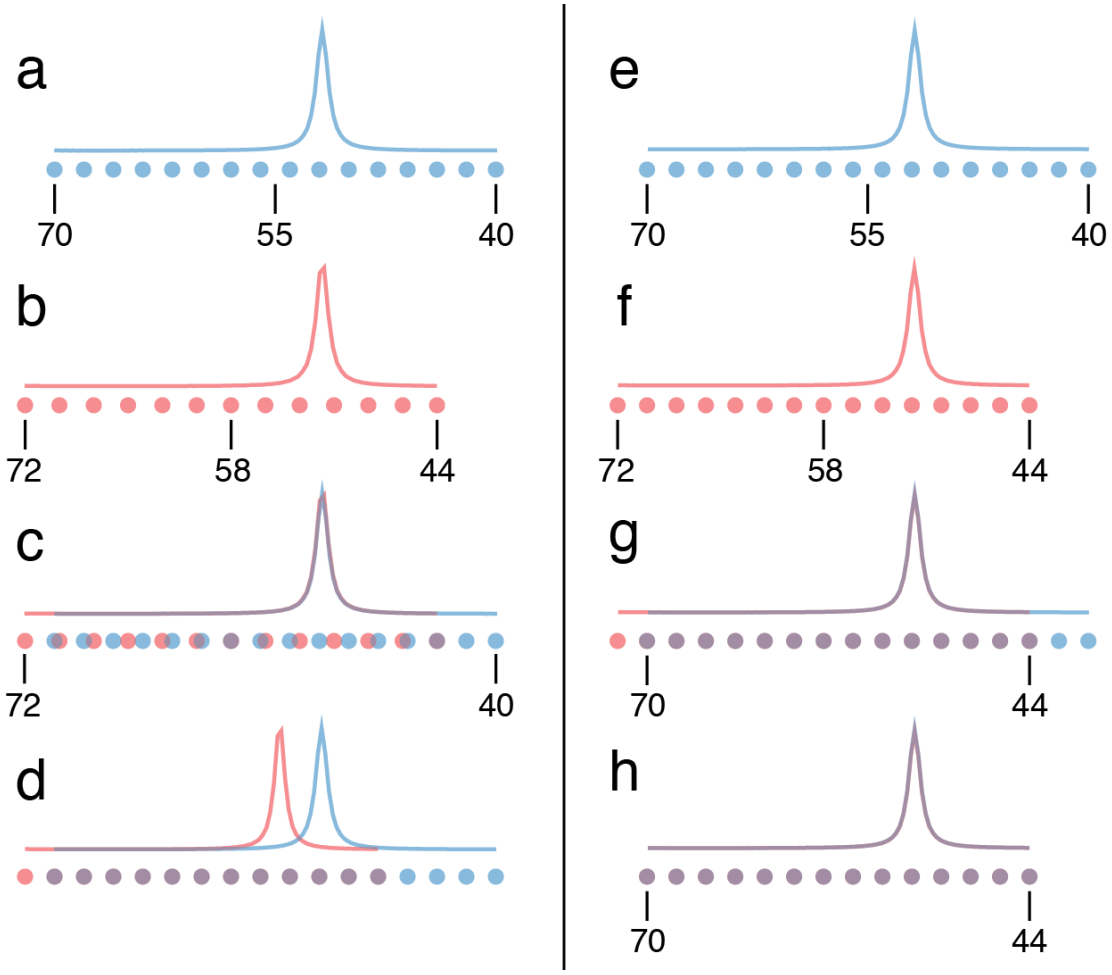

Figure 3.6 Matching the digital resolution between spectra The dots below the spectra correspond to the points at which the frequency axis is sampled. Each space between the dots represents 8 sampled points. (a) A 1D spectrum of 120 points, with a spectral width of 30 PPM and centered at 55 PPM (digital resolution of 4.00 points per PPM). A single signal is present at 52 PPM. (b) A 1D spectrum of 96 points, with a spectral width of 28 PPM and centered at 58 PPM (digital resolution of 3.43 points per PPM). This spectrum also has a signal at 52 PPM. (c) When overlaying the two spectra from (a) and (b), the signals properly align at 52 PPM. However, each spectrum is sampled at very different locations along the frequency axis. (d) If we compress the spectrum from (b) to align with the spacing of those in (a), our signals no longer align. Multiplying the aligned points would not produce a correlation in the covariance spectrum. Consequently, element-wise multiplication is not well defined. (e) The same spectrum as in (a). (f) The spectrum from (b) zero padded from 96 to 112 points before Fourier transformation. It now has a digital resolution of 4.00 as well. (g) When overlaying the signals from (e) and (f), both the signals and the sampling locations align. (h) Each spectrum has been restricted to the region common to both spectra. Element-wise multiplication is well defined and will lead to a correlation in the covariance spectrum.

Finally, users should be sure that all NMRPipe spectra are self-consistently calibrated in every dimension. Combining concepts from Figure 3.4 and Figure 3.6, we note that it is extremely important for each signal to align exactly with its counterpart in another spectrum when undergoing element-wise multiplication. Misalignment due to incorrect calibration, even by one point, can substantially reduce the amplitudes of covariance correlations. We advise users to find a set of clearly identifiable signals and use **nmrDraw** to verify that those signals occur at *exactly* the same chemical shift in every spectrum, usually by updating the values of CAR in **fid.com** scripts.

Once the spectra have been transposed and calibrated and the digital resolutions have been matched, the final consideration before running our script is the size of the resulting 4D spectrum. The size of the 4D spectrum can be calculated by multiplying the total number of points (the product of the dimension sizes I, J, L and M) by the size of each point (4 bytes), and adding the size of the NMRPipe header (2048 bytes). Dividing this number by 2^{20} will give the size in megabytes (MB), whereas dividing it by 2^{30} will give the size in gigabytes (GB). Users should be aware that it may not be possible to manipulate and view large files properly in all software. If users determine that the size will be too large, they should either reduce the size of the input spectra in the I, J, L and/or M dimensions or downsample the resulting 4D spectrum (see below). Reducing the size of the input spectra can be accomplished in three ways: extract only the regions of the spectrum relevant to analysis; reduce the amount of zero padding if any; or

truncate the time-domain data at the cost of resolution. The option to downsample the 4D spectrum will be discussed further below.

3.2.2 Running the script

Users should make a new copy of our script for each calculation. This has the beneficial effect of maintaining a record of processing parameters. Users will need to modify the eight parameters at the beginning of the file: *filenames*, *extract_IJLM*, *extract_K*, *filename_4D*, *labels_4D*, *lambda*, *with_mrs* and *downsample*.

The variable *filenames* is a two-column cell array representing the input spectra. Each row specifies a pair of 3D spectra that will be covaried to produce a 4D spectrum. If multiple rows are present, the 4D spectrum from each row will be combined with all others using element-wise multiplication to form the final spectrum. Each element of the cell array is a string indicating the location of a 3D spectrum. The string may indicate a single file, if it points to a monolithic NMRPipe spectrum, or it may be a standard NMRPipe formatting string of the form '/path/to/test%03d.ft3'. Users should note the order of their spectra in each row. The first column should always indicate the spectrum with dimensions I, J and K, while the second column should always indicate the spectrum with dimensions L, M and K. For example, a correlation map for sequential amide resonance assignment employing HNCA and HN(CO)CA, as well as HN(CA)CB and HN(COCA)CB would require the lines:

```
filenames = {"/path/to/HNCA/ft3/test%03d.ft3", ...
             "/path/to/HNCOCA/ft3/test%03d.ft3"; ...
             "/path/to/HNCACB/ft3/test%03d.ft3", ...
             "/path/to/HNCOCACB/ft3/test%03d.ft3" ...
            }
```

If different spectral widths have been used, the script automatically extracts the largest possible region for each dimension (e.g. 44 PPM to 70 PPM in Figure 3.6). Alternatively, if a specific region is desired, users may specify the limits of the region, in PPM, for each dimension. This is particularly useful when calculating residue-specific covariance maps[44]. A single NMRPipe spectrum can be prepared, and various residue specific regions can be extracted from it during each calculation.

The variables *extract_IJLM* and *extract_K* define the limits of extracted regions for the various dimensions that undergo element-wise multiplication. *extract_IJLM* is a 1 by 4 cell array, where each element is a 1 by 2 array representing the limits of extraction for the I, J, L and M dimensions respectively. Be sure to verify that these limits fall within the bounds of each spectrum. In some cases, the nominal value for the edge of a spectrum may be slightly different from its true value. If an array is empty, the script will default to the largest possible region for that dimension. If no specific regions are desired for any of the four dimensions, users can leave the entire cell array empty. *extract_K* defines the regions extracted along covaried dimension. It functions exactly as *extract_IJLM*, except that it is a 1 by P cell array where P is the number of pairs of 3D spectra to be combined in the final 4D spectrum. This number should also correspond to the

number of rows in the *filenames* cell array. Once again, if no specific regions are desired for any of the P different K dimensions, users can leave the entire cell array empty. The following examples are all valid inputs when P = 3:

```
extract_IJLM = {[6 10.5], [131 103], [10.5 6], [103 131]};
extract_IJLM = {[6 10.5], [], [10.5 6], []};
extract_IJLM = {[], [], [], []};
extract_IJLM = {};

extract_K = {[30 60], [15 75], [182.5 171]};
extract_K = {[], [15 75], []};
extract_K = {[], [], []};
extract_K = {};
```

The variable *filename_4D* represents the output spectrum. The spectrum may be output in one of three NMRPipe formats: a single, monolithic file; a series of 2D files; or a series of 3D files. If a monolithic file is desired, *filename_4D* should be a string specifying a single file, for example ‘path/to/test.ft4’. In some cases, monolithic files may require an additional processing step after our script has run. We managed to avoid this constraint for Mac and Linux but not for Windows. When used in Windows, the script will issue a warning reminding users to perform this task. If a series of 2D files is desired, then the string should be in the familiar NMRPipe format for 4D spectra, for example '/path/to/test_%03d_%03d.ft4'. Finally, if a series of 3D files is desired, the string should be similar to that of a 3D NMRPipe spectrum, for example '/path/to/test%03d.ft4'.

The variable *labels_4D* specifies an NMRPipe label for each dimension in the resulting 4D spectrum. It is a 1 by 4 cell array, where each element is a string of up to four characters. The order of the labels coincides with the order of the dimensions, namely I, J, L and M. For example:

```
labels_4D = {'HN', 'N', 'HM', 'CM'};
```

The variable *lambda* corresponds to the parameter λ in the Generalized Indirect Covariance formalism[78] and reflects the power to which the covariance spectrum is taken during calculation. λ values of 1/2 were used to suppress the effects of pseudo-relay artifacts in symmetrical covariance of NOESY and TOCSY type experiments[68]. In unsymmetrical covariance, the slope of a signal with respect to λ can be used to assess the signal's veracity[78].

The boolean variable *with_mrs* specifies whether maxima ratio scaling should be implemented. Maxima ratio scaling is a technique developed by Snyder and coworkers[79] to reduce the effects of inhomogeneous noise in covariance spectra. Because each point in a covariance spectrum is the product of two 1D slices, the noise at each point is modulated by the amplitude of the signals in the slices from which it came. This has a non-linear effect on the distribution of the noise and can often create "ridges" of noise emanating from strong signals in the covariance spectrum. In a 4D spectrum, instead of ridges, this often manifests as varying noise intensities in each different plane of the spectrum. Turning on maxima ratio scaling can help to reduce this effect.

Finally, *downsample* is a parameter used to reduce the size of the output spectrum. Rather than limiting the size of the input spectra to reduce the size of

the 4D spectrum, possibly broadening its signals, downsampling seeks to use the highest resolution input spectra possible and simply reduce the size of the 4D spectrum after the fact. It is applied to the dimensions I, J, L, and M and not to the subsumed dimension K. Downsampling reduces the number of points in a particular dimension by writing only every n^{th} point along that dimension. The variable *downsample* is a 1 by 4 cell array with integers defining the amount of downsampling in each dimension (I, J, L or M). A downsampling factor of one has no effect on the data. A downsampling factor of two indicates that the script should only output every 2^{nd} point along that dimension. If downsampling is not desired in any dimension, simply leave the cell array empty. The following examples illustrate valid input:

```
downsample = {2, 2, 2, 2};
downsample = {2, 1, 2, 1};
downsample = {1, 1, 1, 1};
downsample = {};
```

While downsampling may seem like a promising solution to reduce data size, we caution users on its use. When signals are relatively sharp in the original data, some correlations may be severely reduced by the procedure. In general, we have found that it is safer to reduce the size of input spectra. Nevertheless, we have retained this option for users to experiment with.

3.3 Navigating and interpreting 4D covariance spectra

Just as a 3D spectrum can be thought of as a series (line) of planes, a 4D spectrum can be thought of as a plane of planes. In a 3D spectrum, we often use

a cursor to choose coordinates in two dimensions and view the corresponding 1D slice along the remaining dimension. In a 4D spectrum, choosing coordinates in two dimensions allows us to view the corresponding 2D plane along the remaining dimensions. A 4D spectrum viewer typically displays two planes of data side by side, each with its own cursor. The plane on the right is specified by the two coordinates from the cursor on the left and vice versa. Moving the cursor in one plane updates the data in the opposite plane based on the new coordinates.

Navigating a 4D spectrum is much like navigating a 3D spectrum. In a 3D spectrum, the two dimensions corresponding to a spin anchor are usually synchronized to an HSQC. For example, when viewing an HNCA, users often use an HN-HSQC to choose a signal of interest and then identify the location of its corresponding C^{α} chemical shift in a synchronized display of an HNCA spectrum. In a 4D spectrum, however, there are two spin anchors that can both be synchronized to HSQCs. Extending the 3D example to a 4D spectrum of (H^{N}, N, H^{α}, C^{α}) correlations, the H^{N} and N dimensions can be synchronized to an HN-HSQC while the H^{α} and C^{α} dimensions can be synchronized to an HC-HSQC. In this scenario, choosing a signal of interest in the HN-HSQC allows users to identify its corresponding signal in the HC-HSQC.

Figure 3.7 illustrates the navigation of a 4D covariance spectrum and demonstrates some of the complications that might arise when assigning resonances. Here, we discuss a situation exposing the limitations of pre- and post-processing artifact suppression, and we present a solution to overcome these limitations during analysis. The situation arises when a weak and a strong signal

have nearly degenerate frequencies in the covaried dimension, and the (reduced) artifact correlation in the 4D covariance spectrum competes with the true, weak correlation. Figure 3.7 displays a cartoon representation of a 4D spectrum generated from HNCA, HN(CA)CB and HMCMCBCA spectra and connecting H^NN spin anchors to H^MC^M spin anchors in valine residues. Panels (a) and (b) display selected regions of the HN- and HC-HSQCs associated with this data; note the differing peak intensities. In the following example, it will be assumed that the H^NN spin anchors are fully assigned and that we are using the 4D covariance spectrum to assign their associated methyl resonances. To emphasize this strategy, we will refer to the H^N/N plane of the 4D as the fixed plane and the H^M/C^M plane as the explored plane. Panels (c-f) represent various states of the 4D spectrum throughout the assignment process.

Panel (c) represents a view of the 4D spectrum with the cursor positions of each plane synchronized with their respective HSQCs as defined in (a) and (b). In this initial configuration, the cursor in the fixed plane is set to investigate a weak signal observed in the HN-HSQC. As a result, we can identify several possible methyl candidates in the explored plane. On the other hand, because we have not set out to investigate a particular methyl signal, the cursor in the HC-HSQC is placed at a position containing only noise. Accordingly, the H^N/N plane has no visible signals.

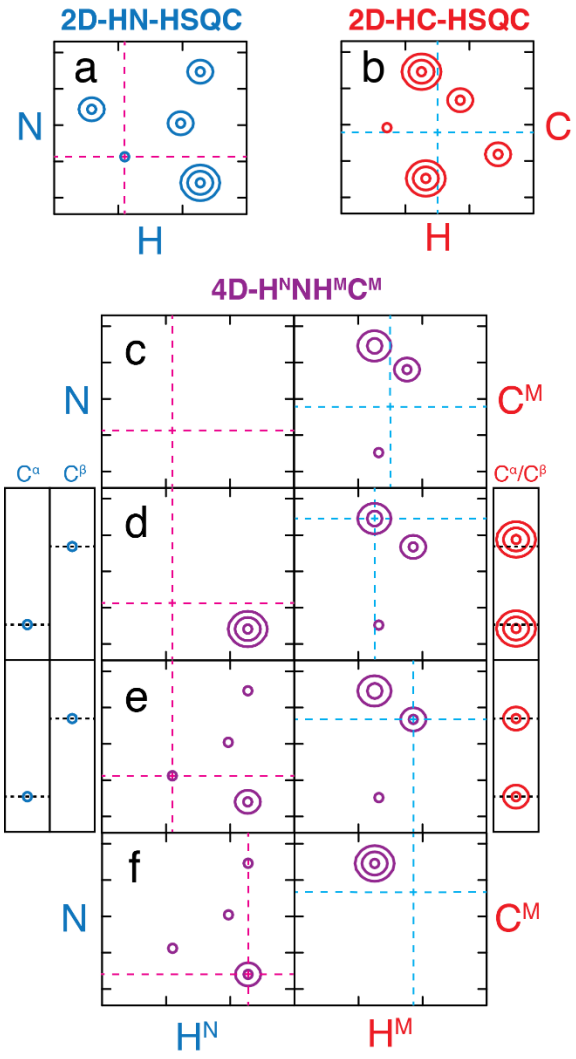

Figure 3.7 Navigating a 4D spectrum (a) Zoomed region of a 2D HN-HSQC. (b) Zoomed region of a 2D HC-HSQC. (c) – (f) Views of a 4D covariance spectrum connecting H^NN spin anchors to H^MC^M spin anchors through C^α and C^β. The left panes represent the H^N/N planes at the position of the H^M/C^M cursor on the right panes. The right panes depict the H^M/C^M planes at the position of the H^N/N cursor on the left panes. The panes to the left of panels (d) and (e) show strips from the HNCA and HN(CA)CB along the C^α and C^β dimensions at the position of the H^N/N cursor. Similarly, the panes to the right show strips from the HMCMCBCA at the position of the H^M/C^M cursor. (c) The cursors begin synchronized to their respective positions in the 2D HSQCs. The right pane shows three possible methyl assignments for the amide signal selected in the HN-HSQC. No peak is selected in the HC-HSQC, so no signals are present in the left pane. (d) One of the assignment candidates is selected in the H^M/C^M plane revealing (H^N, N) correlations in the H^N/N plane but none at the coordinates of the cursor. The positions of signals in the strips do not match. (e) A second assignment candidate is selected in the H^M/C^M plane. Here, the positions of signals in the strips match and the methyl resonances have been assigned to the amide resonances. (f) Moving the H^N/N cursor to a peak discovered in (d) reveals its assignment candidates.

Panel (d) illustrates the 4D spectrum after moving the cursor in the explored plane to its largest signal, updating the view of the fixed plane. We observe a strong amide signal in the fixed plane, but not at the H^N/N position of interest. Because the dynamic range in covariance spectra is larger than that of normally acquired spectra, users should expect to frequently change the contour levels as needed during analysis (as assumed when creating Figure 3.7). In this scenario, the contour levels have been set according to the intense amide signal, and the signal at the (H^N, N) position of interest lies below the contour threshold. This is an immediate indicator that, in all likelihood, this is not the correct methyl assignment for the amide signal under investigation. We can verify this hypothesis by inspecting the original 3D data. The H^N/N cursor of the 4D spectrum is synchronized to those of HNCA and HN(CA)CB and the H^M/C^M cursor is synchronized to that of HMCMCBCA, providing the strips shown to the left and to the right, respectively. Inspection of the carbon dimensions reveals that both C^α and C^β are nearly degenerate but clearly different for H^NN and $H^M C^M$ anchors (horizontal lines). This false-positive correlation occurs in spite of the spectral derivative in the carbon dimensions because of the large amplitude of the signal in the HMCMCBCA spectrum. Because such a scenario cannot be predicted, we use covariance spectra to supplement rather than supplant examination of the original data. That is, we maintain the synchronization with the original data (here HNCA, HN(CA)CB, and HMCMCBCA) throughout the entire assignment procedure.

Panel (e) investigates the next-strongest signal in the explored plane. In addition to the previously discussed signal in the fixed plane, we can now observe

a small signal at the position of interest. However, we also note that there are many such small signals in the fixed plane. When assigning correlations involving a weak resonance correlated with a strong signal in the original spectra, it is often much easier to search for candidates by selecting the dimensions of the weak peak as the "fixed" plane and search for new correlations in the explored plane with dimensions of the strong signal. In effect, the amplitude of the entire explored plane (including noise) reflects the amplitude of the signal in the fixed plane. In our experience, the correct assignment is more likely to stand out when searching in this direction. In the other direction, many false-positive correlations may be of equal or greater height than the true correlation. In this case, examination of the original data reveals identical C^{α} and C^{β} resonance frequencies in the associated strips, and after only investigating two candidate resonances, we have assigned a methyl signal for this valine residue without ever having picked a peak. Simultaneously, in panel (d), we have identified a candidate amide signal for a second methyl. This assignment is confirmed in panel (f), where the cursor in the fixed H^N/N plane has been moved to the amide candidate, revealing a single signal in the H^M/C^M explored plane. Importantly, if the explored plane is devoid of any signals, it indicates conclusively that no such correlation exists in the raw data. This is in contrast to peak picking techniques where it is not possible to distinguish between truly missing signals and unpicked peaks.

3.4 Utility of 4D covariance spectra in large proteins

3.4.1 Examples of 4D covariance spectra

Our script can be used to create 4D spectra connecting any combination of HN or HC spin anchors. Figure 3.8 presents a number of possible combinations. Each pair of 3D spectra to be covaried should feature correlations to a common nucleus that will be used to correlate the two different spin anchors. Further, if the same spin anchors can be related to each other through different nuclei, the resulting set of 4D covariance spectra can be combined with element-wise multiplication to reduce artifacts.

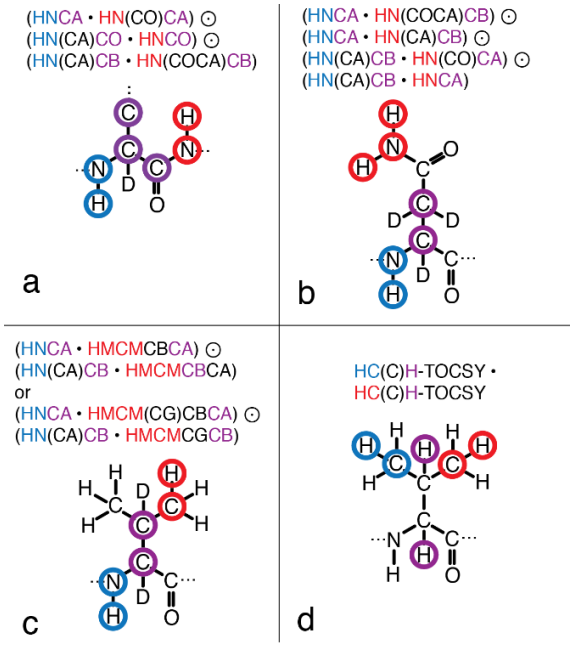

Figure 3.8 Types of 4D covariance spectra Each panel illustrates the nuclei involved in a particular 4D covariance spectrum. The nuclei forming the spin anchors that will appear in the 4D spectra are circled in blue or red, while the nuclei subject to covariance and connecting the two spin anchors are circled in purple. Above each amino acid is a diagram indicating the 3D spectra involved and their proper combination during covariance calculations. A single dot indicates that two spectra should be covaried, subsuming the purple dimensions. A circled dot indicates that two covariance spectra should be multiplied in an element-wise fashion.

Panel (a) of Figure 3.8 illustrates what is perhaps the most useful 4D covariance map. Here, sequential amide H^NN spin anchors are connected to each other through their mutual correlations to C^α, C^β and C' nuclei. This process utilizes the same spectra as traditional strip matching approaches to assign backbone resonances, but it does so without relying on error prone peak lists. The advantages of the covariance approach relative to traditional backbone assignment methods are discussed further in the next section.

In addition to sequential backbone assignment, 4D covariance spectra can also be used to connect backbone H^NN spin anchors to their counterpart sidechain $H^\delta N^\delta$ or H^MC^M spin anchors. Panels (b) and (c) illustrate two such examples relevant to large proteins. In the first case, where H^NN spin anchors are connected to $H^\delta N^\delta$ spin anchors, only the traditional backbone assignment experiments are necessary. The asparagine sidechain mimics the structure of a peptide bond, and as a consequence, C^β nuclei appear as a pseudo-C^α in HNCA and HN(CO)CA experiments while C^α nuclei appear as a pseudo-C^β in HN(CA)CB and HN(COCA)CB experiments. Covarying an HNCA (connecting H^NN spin anchors to C^α) with an HN(CA)CB (connecting $H^\delta N^\delta$ spin anchors to C^α) yields a 4D-$H^NNH^\delta N^\delta$ covariance spectrum (Figure 3.9). Similar combinations can be created with all four of the C^α and C^β spectra discussed above, and the resulting 4D spectra can be combined with element-wise multiplication. Additionally, we note that a similar spectrum is possible for glutamine residues, but in our hands degeneracy at the C^β position limits its usefulness. In the second case, where H^NN spin anchors are connected to H^MC^M spin anchors, additional HMCMCBCA and HMCM(CG)CBCA

experiments are necessary. These experiments connect methyl signals to their counterpart aliphatic signals. For example, in valine residues the HMCMCBCA experiment yields (H^M, C^M, C^α) and (H^M, C^M, C^β) correlations. Covarying this spectrum with both an HNCA and an HN(CA)CB results in two 4D-$H^NNH^MC^M$ spectra (Figure 3.10), which can again be combined with element-wise multiplication. In both cases of sidechain assignment, the covariance spectra convert what is normally a tedious and cumbersome assignment process that requires simultaneous analysis of many different spectra into one of simple inspection of a 2D plane from a 4D spectrum. Figure 3.9 and Figure 3.10 illustrate the success and simplicity of this assignment strategy.

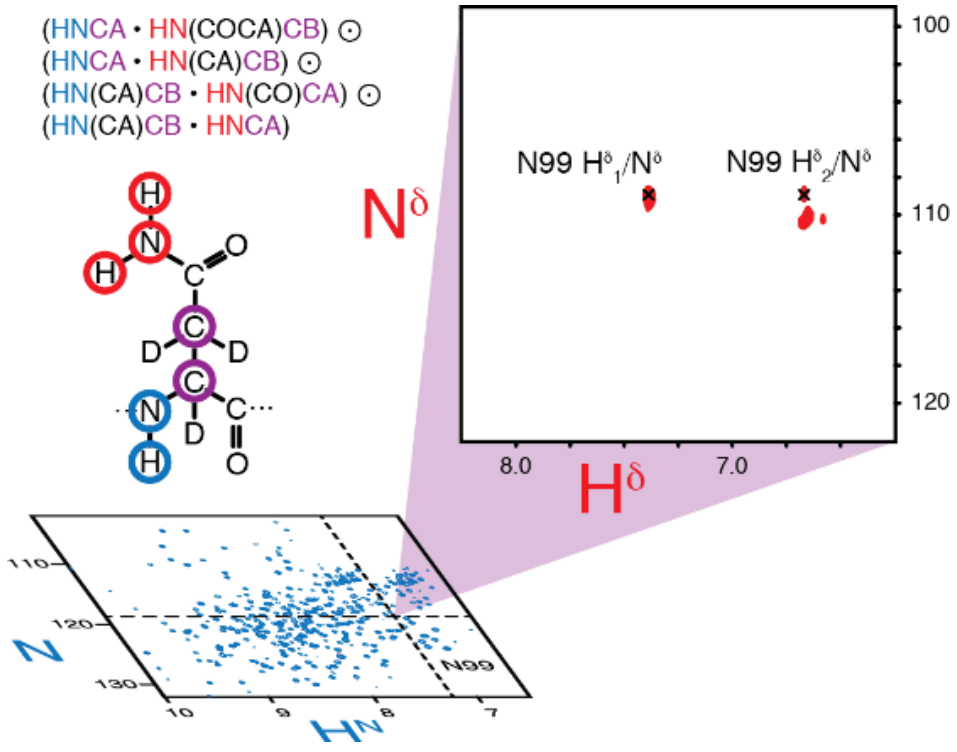

Figure 3.9 Asparagine sidechain assignment in $Cy1_{ybt}$ The blue spectrum, laid flat, is an HN-TROSY experiment. The cursor in this spectrum indicates the position of N99. The red spectrum is an H^δ/N^δ plane taken from the 4D-ASN covariance spectrum at the position of the cursor in the H^N/N plane. It reveals the positions of N99's corresponding sidechain $H^\delta N^\delta$ spin anchors.

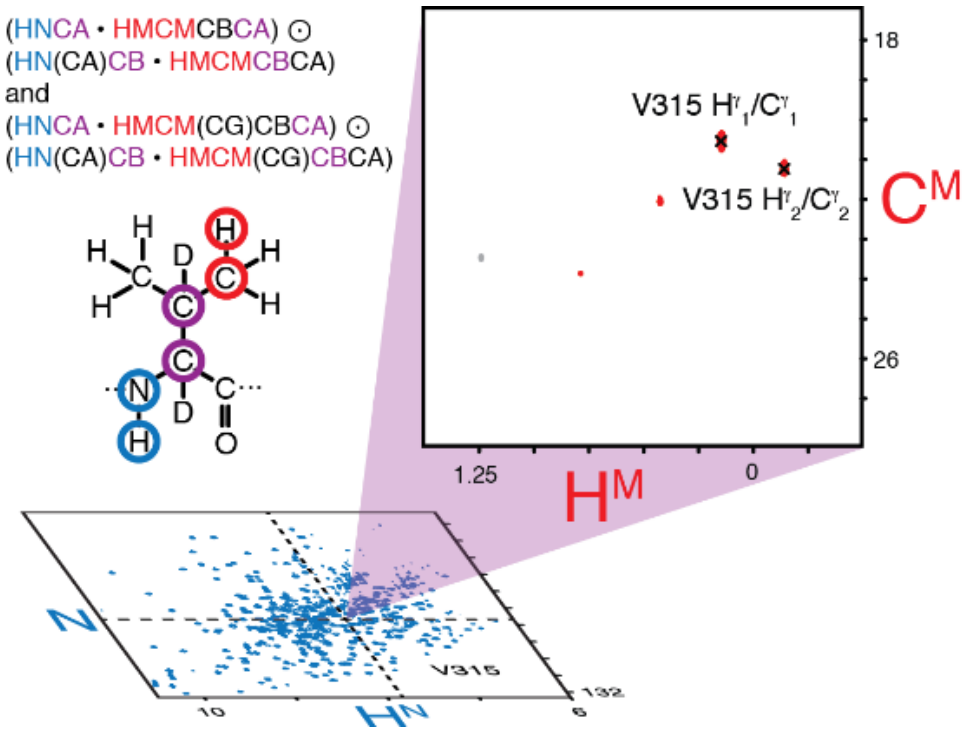

Figure 3.10 Methyl sidechain assignment in $Cy1_{ybt}$ The blue spectrum, laid flat, is an HN-TROSY experiment. The cursor in this spectrum indicates the position of V315. The red spectrum is an H^M/C^M plane taken from the 4D-VAL covariance spectrum at the position of the cursor in the H^N/N plane. It reveals the positions of V315's corresponding sidechain $H^{\gamma}C^{\gamma}$ spin anchors.

The final example in Figure 3.8, panel (d), illustrates a 4D covariance spectrum that connects sidechain aliphatic spin anchors to each other using a 3D-HC(C)H TOCSY spectrum. In a sense, this covariance spectrum acts to reconstruct the 4D-HCCH TOCSY spectrum of which the 3D-HC(C)H TOCSY is a projection. This approach is useful in smaller proteins to help complete the sidechain assignments. An example of such a spectrum is shown in Figure 3.11.

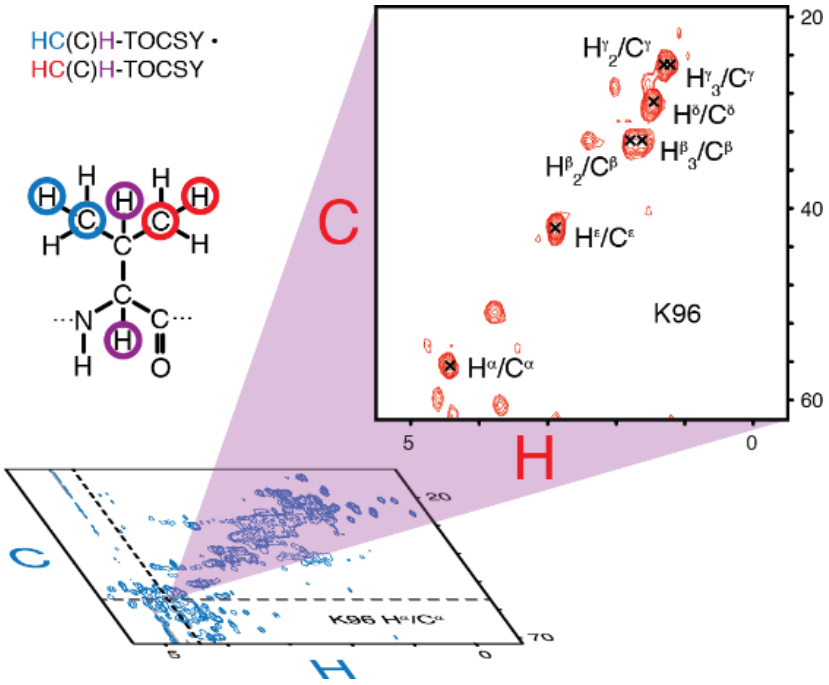

Figure 3.11 Sidechain assignment in PCP1$_{ybt}$ The blue spectrum, laid flat, is an HC-HSQC experiment. The cursor in this spectrum indicates the H$^{\alpha}$C$^{\alpha}$ position of K96. The red spectrum is an H/C plane taken from the 4D-HCCH-TOCSY covariance spectrum at the position of the cursor in the HC-HSQC. It reveals the positions of all of the sidechain HC spin anchors for K96.

3.4.2 Comparison to traditional backbone assignment

Sequential NMR assignment procedures for backbone nuclei seek to identify H^{N}N spin anchors belonging to adjacent residues in the polypeptide chain. This task is accomplished using through-bond correlations to C$^{\alpha}$, C$^{\beta}$, and C' carbon nuclei along the peptide backbone. These (H, N, C) correlations are detected using two classes of 3D experiments. The first provides correlations between H^{N}N spin anchors and carbon nuclei of both the same and the preceding residue, i.e. correlations of the form (H_i, N_i, C_i) and (H_i, N_i, C_{i-1}). We refer to this type of experiment as Intra-3D to emphasize the former correlation. The second type, Seq-3D, provides correlations only with carbons of the preceding residue, i.e. (H_i, N_i, C_{i-1}). Thus, HNCA, HN(CA)CB, and HN(CA)CO are Intra-3D experiments

whereas HN(CO)CA, HN(COCA)CB, and HNCO are Seq-3D experiments. While these experiments, or some subset thereof, are used directly in traditional peak picking and strip matching approaches, they can instead be combined into a single 4D covariance spectrum for analysis. Here the Intra-3D spectrum for each carbon nucleus is covaried with its corresponding Seq-3D spectrum, and the results are combined with element-wise multiplication as discussed previously. The resulting 4D covariance spectrum directly correlates sequentially attached H^NN spin anchors. We have termed such 4D covariance sequential correlation maps 4D-COSCOMs.

The availability of 4D-COSCOMs drastically simplifies assigning the backbone H^N and N nuclei of large proteins. In what follows we discuss the assignment of backbone resonances of three sequential residues from $Cy1_{ybt}$, a 52 kDa cyclization domain from the NRPS Yersiniabactin Synthetase. This stretch exemplifies the advantages of 4D-COSCOMs relative to traditional strip matching.

Figure 3.12 (a, b, d) display three H_{i+1}/N_{i+1} planes taken from a 4D-COSCOM utilizing the C^{α}/C' subset of $Cy1_{ybt}$ backbone experiments. The planes display the sequential dimensions of the 4D taken at three different (H_i, N_i) positions: A415, L416, and V417. The first plane, taken at the (H_i, N_i) position of A415, demonstrates the overall fitness of 4D-COSCOMs (Figure 3.12 (a)). Here we see only two strong peaks: that of the truly sequential residue, L416, and the "auto" correlation at A415. The latter peak occurs when sequential correlations are present in Intra-3D spectra. Such correlations could be eliminated by the use of intraresidual-HNCA and intraresidual-HN(CA)CO experiments[82–85], but in our

hands the associated sensitivity losses have precluded their use. The second H_{i+1}/N_{i+1} plane features three strong signals (panel (b)). The "auto" correlation at L416 and the true sequential correlation at V417 are present. However, there is also a strong peak at the position of K321. Examination of this correlation reveals that its predecessor C^{α}_{i-1} and C'_{i-1} chemical shifts do indeed match the intra C^{α}_{i} and C'_{i} signals of V416. Indeed, the correct successor to V416 could only be identified through careful comparison of the original 3D data sets along with cross-validation using NOESY spectra. This example highlights an advantage of COSCOMs relative to strip matching. Here, the two possibilities are realized directly and immediately, whereas when using strip-matching, the same conclusion could only be reached after the comparison of multiple strips. The third and final H_{i+1}/N_{i+1} plane (panel (d)), taken at the position of V417, contains an "auto" correlation as well as correlations to both E418 and Q430. This phenomenon is the result of ^{1}H and ^{15}N overlap between residues V417 and N429 in the original 3D data. Indeed, Figure 3.12 (c) contains a 2D $^{1}H/^{15}N$ projection of the $Cy1_{ybt}$ HNCO, and its inset demonstrates that V417 overlaps strongly with N429. Finally, the additional correlation to F442 seen in panel (d) highlights a vulnerability of COSCOMs. This correlation is another consequence of sequential correlations in Intra-3D experiments. While the C'_{i-1} of F442 does indeed match the C'_{i} of V417, the C^{α}_{i-1} of F442 matches the sequential C^{α}_{i-1} peak of N429 rather than either of the intra C^{α}_{i} peaks present in the slice. In a similar fashion, lingering sequential correlations in Intra-3D spectra do occasionally create erroneous peaks other than the

expected “auto” correlations. Therefore we emphasize that 4D-COSCOMs are intended to supplement rather than supplant traditional assignment procedures.

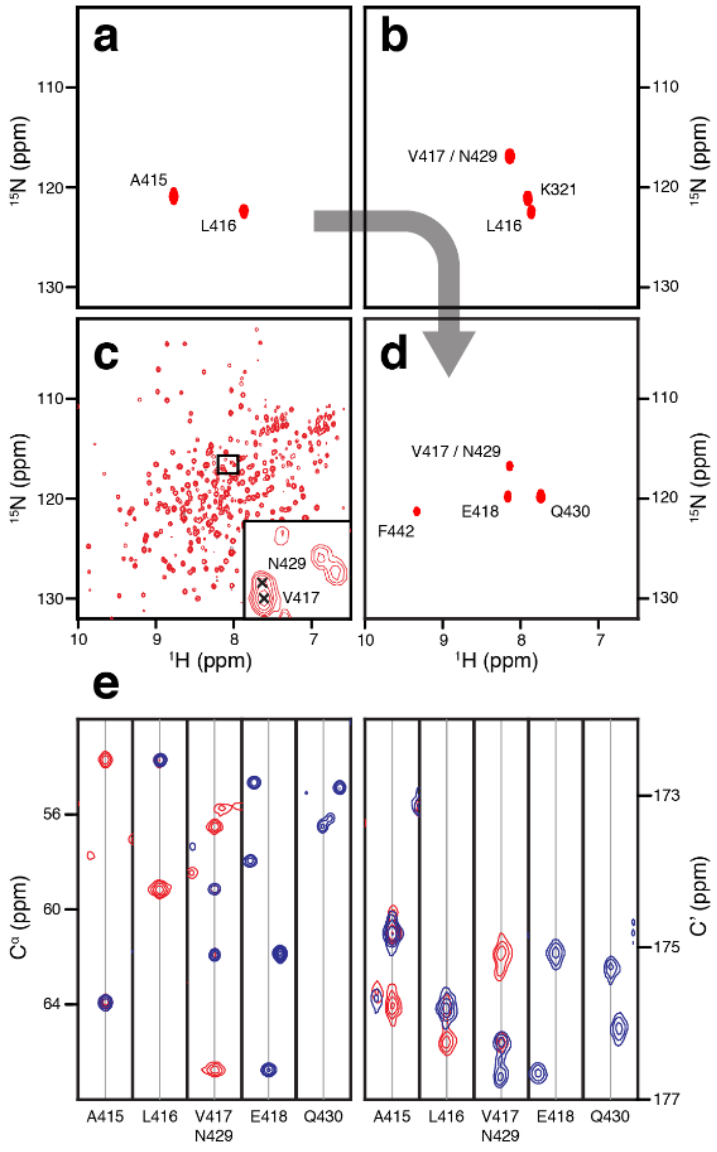

Figure 3.12 4D-COSCOMs of Cy1$_{ybt}$ (a) H_{i+1}/N_{i+1} plane from the 4D at the (H_i,N_i) position of A415 (b) H_{i+1}/N_{i+1} plane at L416 (c) $^{1}H/^{15}N$ projection of the HNCO. The inset shows a zoomed region around the overlapped residues V417 and N429. (d) H_{i+1}/N_{i+1} plane at V417. (e) Traditional strip matching of Cy1$_{ybt}$. Intra-3D spectra are shown in red and Seq-3D spectra are shown in blue. C$^{\alpha}$ spectra are shown to the left and C’ spectra are shown to the right.

Panel (e) contrasts the traditional strip matching approach with that of the 4D-COSCOMs shown above. With strip matching, it is often the case that several

candidate resonances must be investigated side-by side while displaying many carbon dimensions before correctly identifying a residue's successor. In the example we discussed two out of three assignments may have been hindered when using strip matching. First, if the C^{α}_{i-1} signal of V417 had either been erroneously picked or not picked, L416 may have been linked to K321 instead of V417. Second, the C^{α} of V417 may have been paired with the C^{α}_{i-1} of N429 (and similarly for C' resonances) thereby producing an artificial sequence of residues. A careful investigator would have had to probe all possible $C^{\alpha}_{i}/C^{\alpha}_{i-1}$ and C'_{i}/C'_{i-1} combinations to investigate which of them lead to successors. The very nature of COSCOMs overcomes these issues effortlessly; all possible sequential candidates are displayed in an H/N plane. Only those sequential candidates that simultaneously match both C^{α}_{i} and C'_{i} resonances are shown. There is no need to investigate multiple $C^{\alpha}_{i}/C^{\alpha}_{i-1}$ and C'_{i}/C'_{i-1} combinations for V417 and N429; Q430 and E418 are simply seen. Nevertheless, COSCOMs are still vulnerable to erroneous correlations involving sequential signals in Intra-3D experiments, and 4D-COSCOMs must be inspected together with the original data. Overall, 4D-COSCOMs offer an intangible benefit in that they present data from multiple spectra in a more simple and intuitive way than strip matching, resulting in faster and more efficient assignment. Because of these advantages and because COSCOMs do not require any additional data beyond that which is already acquired for traditional assignment, we believe they provide a rather useful tool in facilitating, editing, and proof-reading NMR resonance assignments.

4 Measuring NMR relaxation rates with accordion spectroscopy

Modified portions of this text have been published in the *Journal of Biomolecular NMR*[11]

4.1 Accordion relaxation spectroscopy

4.1.1 Principles of accordion spectroscopy

The accordion method can be used to reduce the dimensionality of NMR relaxation experiments; while traditional methods usually collect an entire series of 2D experiments, the accordion method requires only two measurements.

Relaxation rates are measured by monitoring the decay of NMR signals as a function of the length of a relaxation period t_r, under the influence of either longitudinal or transverse relaxation. During a traditional relaxation experiment, the relaxation delay t_r is incremented independently of the heteronuclear chemical shift evolution period t_1, resulting in a series of 2D spectra. The time evolution, I, of a signal over these two dimensions can be written as

$$I(t_1, t_r) = I_0 e^{(-R^* + i\omega)t_1} e^{-R_i t_r} \quad (4.1)$$

where I_0 is the signal intensity and ω is the heteronuclear chemical shift frequency. R_i is the relaxation rate of interest (either R_1 or R_2) and R^* represents any relaxation occurring during t_1, including contributions from field inhomogeneities. Analyzing relaxation data in this form is straightforward; encoding with t_1 provides signal dispersion along a second dimension, and the relaxation rates are extracted by

fitting an exponential to the signal intensities along a third, distinct dimension spanned by t_r. In contrast, the analysis of accordion data is more complicated because the t_1 and t_r dimensions are no longer distinct.

In an accordion experiment, the relaxation and chemical shift evolution periods are incremented synchronously yet with different time increments Δt_r and Δt_1, respectively. As a result, the t_r and t_1 dimensions are combined to form a single dimension. The proportionality constant between time increments

$$\kappa = \frac{\Delta t_r}{\Delta t_1} \tag{4.2}$$

is termed the accordion factor[17]. If we express t_r as a function of t_1

$$t_r = \kappa t_1 \tag{4.3}$$

and rewrite the apparent time evolution of the signal, now as a function of t_1 *only*,

$$I(t_1) = I_0 e^{(-R^* + i\omega)t_1} e^{-\kappa R_i t_1} = I_0 e^{(-R^* + \kappa R_i)t_1} e^{-i\omega t_1} \tag{4.4}$$

we recover an easily interpretable expression. The frequency dependence of this signal is identical to that found in the traditional experiment, leaving the position of peaks in the spectrum unchanged. The signal decay envelope, however, now reports on relaxation during both the t_r *and* t_1 periods, as does the line-shape of the signal in the corresponding frequency domain. We define R_{obs}, the *observed* decay constant in the indirect dimension, as

$$R_{obs} = R^* + \kappa R_i \tag{4.5}$$

Thus, two measurements of R_{obs} with different values of κ enable discrimination between the relaxation rate during t_1 (R^*) and the relaxation rate during t_r (R_i).

Although some protocols set constraints on the values of κ in order to cleverly simplify the analysis[13], SARA has been designed to analyze pairs of spectra with any two different values of κ. Additionally, SARA does not impose any experimental constraints other than requiring that the initial amplitude in t_1 be identical for both experiments. In other words, it only requires that the user not change parameters such as the gain or the number of scans between the two experiments.

Overall, the principle of the accordion method is rather simple, yet the experimental time-savings are substantial: Relaxation rates are encoded into the decay of the indirect dimension (or equivalently into the corresponding signal line-shape after Fourier transformation (FT)) and only two experiments are needed to distinguish R_i from R^*. However, the process of extracting relaxation rates using non-linear fitting procedures, particularly in the context of crowded protein spectra, is a rather complex task best performed with the assistance of interactive software.

4.1.2 Overview of fitting protocols

Many protocols have been suggested for extracting rate constants from accordion spectra. In their original accordion publications[12,86] studying exchange processes, Bodenhausen and Ernst proposed obtaining exchange rates from direct fits of Lorentzians to line-shapes in the frequency domain. They also suggested extracting individual peaks from spectra and monitoring cross-peak buildup for second order exchange processes following inverse Fourier transformation. Mandel and Palmer[13] proposed fitting decaying oscillators to accordion interferograms (i.e. Fourier transformed in the directly detected ^1H

dimension but not in the indirect dimension). Rabier, Lefèvre and coworkers[14] also fit using decaying oscillators, but they computed the residuals following Fourier transformation. Guenneugues et al.[17] fit the line-shapes of signals in the frequency domain. Finally Tjandra and co-workers[15] measured the line-widths of signals to extract relaxation rates. We have found that, in their original implementation, specific constraints limit each protocol (see below), and they may not be applicable individually and in all circumstances. SARA gives users access to a selection of protocols spanning the breadth of those proposed by other groups in order to allow researchers to choose the optimal fitting procedures for a given protein. For each protocol, modifications have been implemented to enable a more general application. SARA also uses a graphical and user-friendly MATLAB interface. In this section, we discuss the details of each fitting procedure.

SARA offers the flexibility to analyze data originating from several processing software packages, but some aspects of processing must conform to a few specific guidelines. Spectra may be loaded in the NMRPipe[48] format or in a format adapted from NMRLAB[87]. To allow for comparison among various procedures, SARA requires two input files: a t_1 interferogram and a fully processed two-dimensional spectrum. There are no restrictions on the processing of the detected dimension (F_2), as this dimension is not part of the relaxation analysis. In contrast, no processing may be applied to the indirect time domain (t_1) of the t_1-interferogram. In the two-dimensional spectrum, only apodization, zero-filling, Fourier transformation, and phasing are permissible when transforming to the indirect frequency dimension (F_1). For convenience, we have provided NMRPipe

and NMRLAB processing scripts specifically designed to output both files in a format compatible with SARA.

SARA also requires a few auxiliary files and parameters for analysis: the value of κ for each spectrum, an X-easy peak list, a file containing the amino acid sequence in single-letter code, and the values of the maximum chemical shift in each dimension (F_1 and F_2) for referencing. The latter values may be verified within SARA using a calibration dialog. The calibration dialog displays the spectrum and overlays the positions of peaks found in the X-easy peak list. The user may interactively adjust the maximum chemical shift in each dimension to ensure proper alignment. In order to perform Monte Carlo simulations the user must also specify a region of the interferogram from which SARA may calculate the standard deviation of the noise. This is also accomplished with an interactive dialog.

After loading their data, researchers may choose from three procedures, which are described in detail below. The first two procedures are variations of the protocol proposed by Mandel and Palmer (MP)[13]. The first is a modified, fully-automated version of the MP method. The second is a user-assisted version of the automated MP protocol and allows the researcher to optimize the fitting parameters returned by the stand-alone MP algorithm. The last procedure, named FT/IFT, is a user-assisted, semi-automated protocol which extracts signals from the spectrum in the frequency domain and analyzes them after inverse Fourier transform (IFT). It addresses the inability of the MP method to isolate the fitting of one peak from that of other peaks in the same t_1 slice.

4.2 MP protocols

Mandel and Palmer proposed two different data analysis schemes, both of which analyze the data in the time-domain of the indirect dimension, and each requires two experiments with accordion factors $\kappa = +\kappa_0$ (forward experiment) and $\kappa = -\kappa_0$ (reverse experiment). In the first scheme, called the "forward-reverse" method, the observed relaxation rate, R_{obs}, is fit separately in each experiment. The relaxation rate of interest, R_i, is calculated using

$$R_i = \frac{R_f + R_r}{2\kappa_0} \tag{4.6}$$

where R_f is the observed relaxation rate in the forward experiment and R_r is the observed relaxation rate in the reverse experiment. In the second analysis scheme, termed the "negative-time" method, the reverse experiment is inverted in time and concatenated with the forward experiment, forming a single data set. In this case, the entire combined data set is fit at once and provides R_i and R^* directly.

Although the two analysis schemes differ in their approach to obtaining R_i, fitting of the interferogram is accomplished similarly in both. The detected ^1H dimension is Fourier transformed to form a t_1 interferogram, and a t_1 free induction decay (FID) is analyzed for each signal. The positions of the t_1 FIDs in the ^1H dimension are determined by inspecting the signals in the 2D Fourier transformed spectrum. Next, Hankel singular value decomposition (HSVD)[88] of each t_1 FID is used to populate a list of time-domain signals putatively present in the slice. HSVD parameterizes these signals by estimating their amplitudes, phases, frequencies and decay rates. We term a collection of these four parameters an oscillator.

Putative oscillators are screened by comparison to the frequencies and phases of signals that are plausibly present in the FID (given their ^{1}H frequency), and erroneous oscillators are removed. The parameters of the surviving oscillators are used as the starting point for a non-linear least-squares optimization. Both the "forward-reverse" and "negative-time" analysis methods model the FID as a sum of damped oscillators, but the target function differs slightly in each.

Although all accordion publications following those by the Palmer group calculate R_i from two separate fits of R_{obs} (i.e. using the same strategy found in the MP forward-reverse scheme), there is an advantage inherent to the MP negative-time technique that has not yet been utilized by other authors. Because the two accordion experiments are combined prior to fitting, R_i is fit directly by fitting both experiments *simultaneously*. This increases the fit precision and bypasses subsequent calculation of R_i and associated error propagation. However, these advantages are a result of global fitting, of which the negative-time technique is a special case, and we note that there is no inherent need to record experiments with the specific values $\kappa = +\kappa_0$ and $\kappa = -\kappa_0$. A similar advantage is obtained with two arbitrary values of κ if the experiments are fit simultaneously using a suitably constructed global non-linear optimization[89]. SARA harnesses this strategy in its implementation of the MP method.

4.2.1 Automated MP protocol

Our implementation of the MP protocol incorporates two approaches. The first is a fully automated algorithm following the protocol outlined in Figure 4.1 and described in this section. The user only needs to specify a handful of tolerances

prior to initiation. The second is a user-assisted version of the same algorithm, in which the user can modify parameters in each processing block of Figure 4.1 before passing the results to the next block. The latter method is discussed in the next section.

The processing blocks themselves proceed according to the MP method as described above and can be regrouped into three steps as follows: *Step 1* (Figure 4.1 blocks 1-4), selection of a t_1 FID; estimation of the number of signals; HSVD; and assessment of the HSVD results; *Step 2* (Figure 4.1 R_{obs} circle), non-linear least squares fit of the individual FIDs; and *Step 3* (Figure 4.1 global fit block), global non-linear optimization of both FIDs. A Monte Carlo error analysis can be performed once global fitting has been completed. If desired, this Monte Carlo analysis can be applied to a user-defined subset of residues. The two accordion data sets (with different values of κ) are treated independently until global non-linear fitting, Step 3, where they are analyzed simultaneously to extract R_i directly.

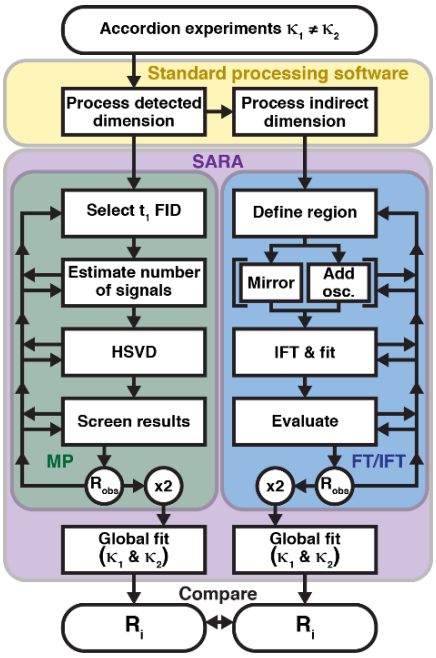

Figure 4.1 Accordion relaxation analysis flow chart. Two accordion experiments are processed with conventional NMR processing software (e.g. NMRPipe) and are loaded into SARA. The interferogram is analyzed by the MP method, while the spectrum is analyzed by the FT/IFT method. In the automated MP method, SARA completes the entire MP analysis. When using interactive methods, users follow the steps of the analysis as outlined in the text. Arrows indicate the flow of data analysis from one task to another. Users may return to any prior step in the analysis at any point in time. Values for R_{obs} must be obtained for each of the two accordion experiments prior to global fitting. Global fitting is accessed via dialogs separate from those of the analysis methods themselves. Results may be plotted in SARA and accessed directly via the SARA save file.

Step 1: Initial estimation of decaying oscillators with HSVD.

The first step selects the appropriate t_1 FID and estimates the number of signals within it. The outcome is a series of amplitudes, phases, frequencies and decay rates that can be used as initial estimates for the non-linear fitting of the FID in Step 2. The FIDs in t_1 are selected with help from the X-easy peak list provided by the user. The user is responsible for verifying that the coordinates specified in the peak list correspond to signal maxima. The peak list is also used to estimate the number of signals present in the FID, a parameter required by HSVD. This number is obtained by counting all signals with proton chemical shifts falling within

a user-defined range of the signal under analysis. Next, HSVD is performed and returns a list of amplitudes, phases, frequencies and decay rates defining the decaying oscillators present in each t_1-FID. HSVD provides reliable estimates of the parameters yet tends to produce spurious oscillators, in particular for weak signals. To resolve this issue, we implement a strategy similar to that employed by Mandel and Palmer. First, we systematically increase the estimate of the number of oscillators provided to HSVD, ensuring that, while erroneous oscillators may be present, all oscillators corresponding to real signals are taken into account. Next, the frequencies returned by HSVD are filtered by comparison to those found in the X-easy peak list, again within the user-defined range of the signal of interest. Thus, in the end, our HSVD procedure produces a total of N oscillators, each with an estimated amplitude (A_n), phase (ϕ_n), frequency (ω_n), and observed decay rate ($R_{obs,n}$) for all N signals present in the FID.

Step 2: Non-linear fit of the individual FID

The next step uses the parameters returned by HSVD as the starting point for a non-linear least squares fit of the FID. The FID is fit to a sum of damped oscillators defined by the target function F(t)

$$F(t) = \sum_{n=1}^{N} A_n e^{i\phi_n} e^{(-R_{obs\,n} + i\omega_n)t} \tag{4.7}$$

The optimized parameter χ^2 is defined as

$$\chi^2 = \sum_{k=1}^{K} Real\{I(k\Delta t_1) - F(k\Delta t_1)\}^2 + Imag\{I(k\Delta t_1) - F(k\Delta t_1)\}^2 \tag{4.8}$$

where I represents the experimental data and k iterates over the points in the discrete time dimension t_1 (i.e. K is the number of points acquired in t_1). This step further refines the parameter estimates from HSVD prior to global fitting. Steps 1 and 2 must be performed on both accordion experiments before beginning Step 3.

Step 3: Perform global non-linear least squares optimization

While HSVD and the preliminary non-linear fits are performed on the FIDs from each accordion experiment separately, the global non-linear least squares optimization is performed on both FIDs simultaneously. The N signals identified in Step 1 and fit in Step 2 are each associated with four parameters: amplitude (A_n), phase (ϕ_n), frequency (ω_n), and observed decay rate ($R_{obs,n}$). Global fitting, however, parameterizes relaxation using the t_r relaxation rate ($R_{i,n}$) and the t_1 relaxation rate (R^*_n). For each oscillator, n, the initial values of R_i and R^* are estimated from the values of R_{obs} using the equations

$$R_i = \frac{R_{obs\,2} - R_{obs\,1}}{\kappa_2 - \kappa_1} \tag{4.9}$$

and

$$R^* = R_{obs\,1} - \kappa_1 R_i \tag{4.10}$$

where the indices 1 and 2 correspond to the two different accordion experiments. The phase of each signal is calculated during each iteration as

$$\phi_n = \omega_n \delta \Delta t_1 \tag{4.11}$$

where δ parameterizes the first evolution time sampled in the experiment, $\delta\Delta t_1$. For most experiments δ = 0 or ½. Initial estimates of the amplitude and frequency for each peak are calculated by averaging the respective values returned by the

individual fits of the two accordion FIDs in Step 2. For this reason it is important that the two experiments are recorded in a manner preserving this amplitude (same gain, number of transients, recycling delay, etc.). This condition is easily verified by comparing the first point in the interferogram, i.e. the 1D-proton spectrum corresponding to t_1=0. Any difference in amplitude between the two experiments will prevent the optimization from converging to the correct values. The fitted function $F_p(t)$ (p = 1, 2) for each of the two FIDs is a sum of N oscillators:

$$F_p(t) = \sum_{n=1}^{N} A_n e^{i\phi_n} e^{(-R_n^* + i\omega_n)t} e^{-\kappa_p R_{i\,n} t} \tag{4.12}$$

The residual to be minimized, χ^2, includes the sum of the squared deviations from both FIDs:

$$\chi^2 = \sum_{p=1}^{2} \sum_{k=1}^{K} Real\{I_p(k\Delta t_1) - F_p(k\Delta t_1)\}^2 + Imag\{I_p(k\Delta t_1) - F_p(k\Delta t_1)\}^2 \tag{4.13}$$

where I and k are defined as in Step 2 and p iterates over the two accordion experiments. Thus, the values of R_i and R^* in $F_p(t)$ are obtained by a simultaneous fit of both experiments. As discussed, including both FIDs in a single fit rather than calculating R_i from individual fits of R_{obs} increases the precision in R_i.

Step 4 – Monte Carlo error analysis

To provide a reliable error estimate, SARA includes a Monte-Carlo error analysis dialog for each method. Prior to its use, however, the level of noise in each interferogram must be established using the noise dialog accessed from the

main SARA menu. Users are asked to define a region in one of the interferograms that contains only noise. Histograms of the real and imaginary values of points within the region are displayed to assist users in identifying outlier regions or artifacts. SARA then calculates the standard deviation of the points in this region for both interferograms with different values of κ to determine the t_1 noise level in each (σ_{n1} and σ_{n2}).

Within the Monte Carlo dialog, the user may specify the residues for which the Monte Carlo will be performed as well as the number of Monte Carlo iterations, N_{MC}. During each round, real and imaginary traces of pseudo-random, Gaussian noise with standard deviation σ_{n1} and σ_{n2} respectively are added to the experimental data of each interferogram. Global fitting, as in Step 3, is repeated N_{MC} times. The output is an estimate of the standard deviations of both R^* and R_i. Once a Monte Carlo error analysis has been performed, error bars are automatically added when plotting results using SARA.

4.2.2 Interactive MP protocol

The interactive MP method follows the same organization as the automated MP method depicted in Figure 4.1 but allows for user intervention at each of the individual blocks. Figure 4.2 shows the window used to optimize the fitting parameters and monitor the fitting quality. As in the automated method, a t_1 FID (displayed at the top left of Figure 4.2) is selected from the interferogram based on the position of a residue in the X-easy peak list. This position is visualized by a blue crosshair cursor in the 2D contour plot (top right, Figure 4.2). The bottom left plot contains the F_1 slice from the 2D spectrum corresponding to the vertical

component of the crosshair and is the Fourier transform of the t_1 FID displayed above it. Displayed at the bottom right is a slice from the F_2 dimension corresponding to the horizontal component of the crosshair in the contour plot. The residue of interest is termed the "active" residue, and the t_1 FID should be chosen to coincide with the maximum intensity of the active residue along the detected frequency dimension (F_2), ensuring the maximum signal-to-noise ratio (SNR) possible. The F_1 and F_2 slices help the user visualize the t_1 FID selection process and are updated interactively upon changes by the user. The user positions the crosshair using the movement controls to the right of the contour plot (“Move” in Figure 4.2). The additional green vertical lines displayed on the 2D contour plot represent the maximum distance in F_2 at which signals are to be considered during HSVD (as in Step 1 of the automated MP procedure). These signals are defined by the "Nearby" button in Figure 4.2, and the estimated number of oscillators present in the FID is updated in an editable box ("# Osc" in Figure 4.2). Vertical lines are added to the F_1 plot (bottom left) at the F_1 position of each peak identified in the search (not shown in Figure 4.2). The vertical line corresponding to the active residue in this plot is marked in green (shown in the F_1 plot of Figure 4.2). These lines allow the user to verify the validity of the estimated number of oscillators. Peaks may appear "nearby" according to the definition implied by the contour plot’s vertical lines but may not appear substantially in the spectrum corresponding to the t_1 FID. With this knowledge the user may adjust the number of oscillators prior to HSVD. Users are reminded that it is usually preferable to over-estimate the number of oscillators, as erroneous oscillators may be removed after HSVD.

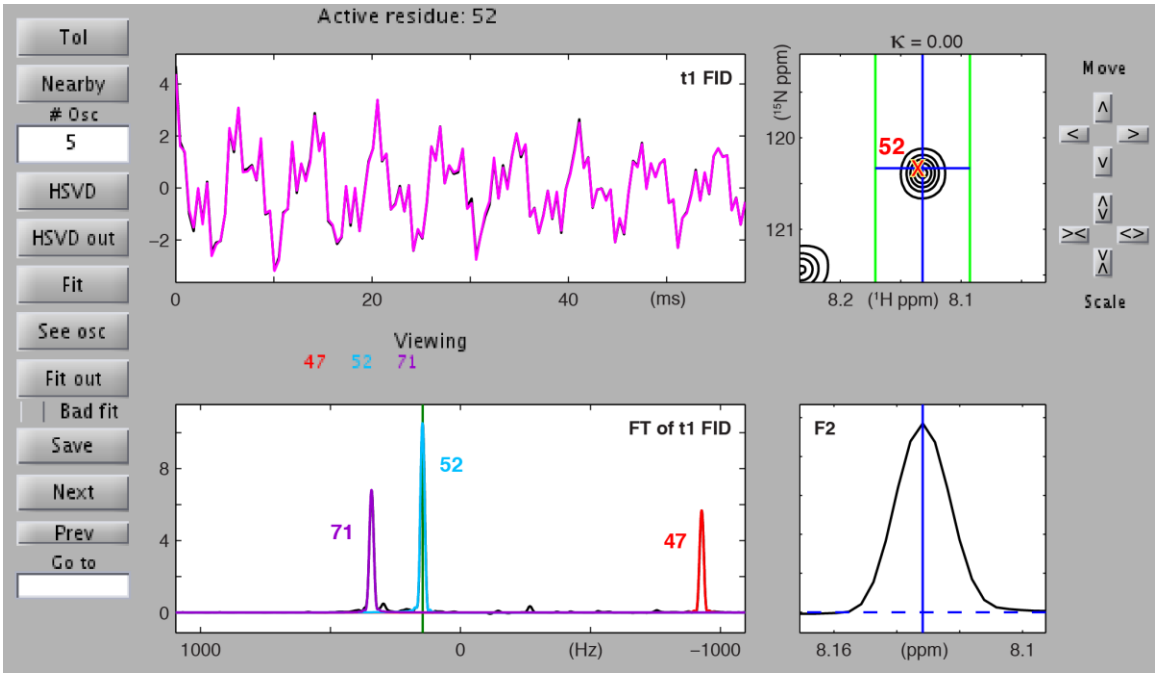

Figure 4.2 The interactive MP dialog White boxes represent user-editable fields. Labels have been edited for clarity. Plots, from left to right and top to bottom, contain: the t_1 FID, a contour plot of the 2D spectrum, the F_1 slice from the 2D spectrum corresponding to the t_1 FID and a slice from the F_2 dimension corresponding to the horizontal line of the blue crosshair in the contour plot. The value of the accordion factor (κ) for the current experiment is displayed above the contour plot (here, "k = 0.00"). Users select the appropriate t_1 FID, with reference to the 2D spectrum, using the "Move" and "Scale" buttons. Nearby residues are identified by pressing the "Nearby" button. The number of oscillators requested from HSVD can be edited ("# Osc"), and the parameters for screening the results can be adjusted ("Tol"). The oscillators returned by HSVD may be edited ("HSVD out") and oscillators may be added or removed. Non-linear fitting is performed with the "Fit" button. The t_1 FID is fit and the resulting fitted curve is displayed in magenta on top of the black experimental data for comparison. The fitted curve in the time domain is Fourier transformed and overlaid on top of the spectrum slice below (not shown), and a green vertical line marks the F_1 position of the active residue. Rather than display the sum of all oscillators in the frequency domain, users may opt to display oscillators individually using the "See osc" feature (red, cyan & purple peaks). If oscillators are viewed individually, their matching residue number is displayed under "Viewing". Fitted parameters are viewed with the "Fit out" button, and poor fits may be marked with the "Bad fit" checkbox. Results must be saved ("Save") or discarded before the user may navigate through the residues of the X-easy peak list ("Next," "Prev," and "Go to").

HSVD of the active residue FID is achieved by pressing the "HSVD" button. As in the automated MP method, the oscillators returned by HSVD are screened by comparison to the expected frequencies and phases of the signals identified in the search for nearby residues. The screening tolerances may be adjusted in the

"Tolerance" dialog ("Tol" button in Figure 4.2). The frequencies of oscillators that pass the screening process are again visualized with vertical lines in the F_1 slice (not shown). Users should verify that there is a one-to-one correspondence between the peaks present in the F_1 plot and the oscillators marked by vertical lines. In the absence of any external information, HSVD provides estimates that are used as the initial values for the subsequent non-linear fit. If better estimates of these parameters have been obtained by other, independent means, the estimates provided by HSVD can be replaced prior to fitting. Additionally, users may specify that a parameter or group of parameters be held fixed during fitting, thus reducing the total number of parameters fitted. Furthermore, entire oscillators may also be added or removed from the fit. All of these changes may be made using the "HSVD out" button in Figure 4.2. Such flexibility when handling fitted parameters enables users to correct inaccuracies found in the MP method, while still benefiting from its high precision.

Until this point, the steps taken by the user have all fallen within Step 1 of the automated MP method and correspond to the blocks in the MP method of Figure 4.1. The user may return to any previous step at any point in time if the parameters are not satisfactory. After editing the results of HSVD, the user presses the "Fit" button (Figure 4.2) to initiate a non-linear fit as described in Step 2 of the automated MP method. The fitted data is overlaid with the experimental data in the t_1 FID plot and is Fourier transformed and overlaid with the F_1 slice below. Users may evaluate the fit using two different methods. The fitted parameters can be examined using the "Fit out" button. Alternatively, the user may display a subset of

the fitted oscillators in the F_1 spectrum, rather than the entire sum, using the “See individual oscillators” feature (“See osc” in Figure 4.2). Together, these methods allow the user to verify that each signal is fit appropriately and that the optimization has not converged to a local minimum. Once the fit is satisfactory, the user must save the results with the “Save” button before either proceeding to another residue in the same experiment (“Next”, “Prev” or “Go To”) or switching to the second accordion experiment. Global fitting (Step 3 of the automated method) may only be performed once the same residue has been successfully fit in both accordion experiments. Both global fitting and Monte Carlo error analysis are performed as described in the automated MP method and are accessed as separate features in the main SARA window.

4.3 FT/IFT protocol

Although we have found the MP method to be reliable in general, severe errors may occur for special cases in crowded spectra. The combination of partial overlap and a high number of oscillators may result in a fit converging to a wrong solution (see Figure 4.4 and section 4.4.1). To help resolve such errors without resorting to comparison with the literature or traditional experiments (e.g. CPMG), which would defeat the purpose of performing the accordion measurement in the first place, we have designed an alternative analysis protocol which we call the Fourier transform/inverse Fourier transform (FT/IFT) method.

The FT/IFT protocol seeks to simplify analysis by fitting signals on an individual basis wherever possible. It is greatly inspired by the original procedure

of Bodenhausen et al., in which the accordion data is Fourier transformed in both dimensions, an F_1 slice is chosen from the 2D spectrum, a peak within the slice is isolated by zeroing all points around it, the slice is inverse Fourier transformed, and finally the magnitude of the resulting time-domain FID is analyzed. Clearly this procedure can only be applied to symmetric, non-overlapped signals. To generalize the method, we have designed a procedure to extract spectral regions featuring one or potentially a few overlapping signals along F_1 while still allowing for subsequent IFT and time-domain fitting. Two solutions are presented to overcome overlap in the extracted region. The first is a simple non-linear fit of the reduced set of overlapping signals in the extracted region. The second involves mirroring half of the line-shape of a partially overlapped signal, a process we call symmetrization, resulting in a single symmetric peak which is then inverse Fourier transformed. These two solutions are not implemented as distinct procedures but rather as alternative features that can be used to assess the reliability of the rates obtained by each method.

The Fourier-transform-based methods presented here allow investigators to fit individual peaks separately. Whereas in the interferogram-based methods proposed by Mandel and Palmer, all signals in the t_1 slice must be fit simultaneously, in the FT/IFT method, signals may be extracted from F_1 slices and fit independently. Simplifying the spectrum to reduce the number of oscillators may offer significant advantages in cases of large proteins or crowded spectra.

4.3.1 Interactive FT/IFT protocol

The FT/IFT method requires two 2D-Fourier-transformed accordion spectra, each with a different value of κ. The steps of the FT/IFT method are summarized in Figure 4.1. Only apodization, zero-filling, Fourier transformation and phasing along the *indirect* dimension are permissible. For each peak, the user defines an extraction region, isolating the signal or group of signals if there is overlap. This region may differ between the two spectra as needed (e.g. to avoid truncation of signals by the region boundaries). Next, we construct a pseudo F_1 slice, placing the extracted region in the center and zero-filling both sides to the full spectral width. Following inverse Fourier transform, the resulting time-domain FID is truncated to the number of points acquired, thus accounting for any initial zero-filling applied during spectrum processing. We refer to this FID as a "reconstructed FID," and it represents a recreation of what would have been acquired if the peak (or group of overlapped peaks) had been isolated and nearly on-resonance during acquisition.

If the extraction region contains only a single, isolated signal then the reconstructed FID is fit as a damped oscillator in a non-linear least squares optimization, bypassing block 2 in Figure 4.1. If, however, the region contains two or more signals that cannot be effectively isolated, then the reconstructed FID can be analyzed in two different manners, as illustrated in Figure 4.1. The first and simplest method models the reconstructed FID as a sum of damped oscillators. The second method requires further spectral manipulation prior to inverse Fourier transform. If a signal is partially resolved from others along F_1, to the extent that

half of its line-shape does not contain any contribution from overlapping peaks, the F_1-slice can be mirrored around the center of the targeted peak in a process we call symmetrization. This procedure results in a single, symmetric signal that can be inverse Fourier transformed and fit using one damped oscillator. In both cases, the fitting procedure is first optimized on fits of R_{obs}, i.e. in reconstructed FIDs derived from each accordion spectrum. The final value of R_i, however, is derived by globally fitting the two reconstructed FIDs simultaneously.

The FT/IFT procedure is semi-automated. The user is required to define the extraction region position and boundaries, establish the number of signals present within the region (usually one), and verify that the non-linear optimization is able to converge. The process is discussed in detail below.

Step 1 – Defining the extraction region

In the FT/IFT method the user navigates through the X-easy peak list and chooses a so-called “active residue.” The X-easy peak list contains the position of each signal in the F_1 and F_2 dimensions along with the signal’s corresponding residue number. This number may be arbitrary in the case of unassigned proteins, but SARA uses it as a label to differentiate signals. When establishing and testing fitting procedures, the active residue is the only residue for which the user should be concerned. While other residues may be included in the fit because of overlap, their rates will be measured from the regions in which each is the active residue. This strategy ensures that the rate measured for each residue is derived using the most optimal extraction region possible.

In SARA an extraction region is parameterized by: an active residue, a list of other residues contained within the region, boundaries of the region in F_1, boundaries in F_2, a 1D slice in F_1, and a 1D slice in F_2. The active residue is listed at the top left of the FT/IFT dialog (Figure 4.3), and any other residues that are to be considered during the fit are listed at the top right. The plots along the top row, from left to right, are: a 1D slice along F_2, a 1D slice along F_1, and a two-dimensional contour plot. The green box in the contour plot represents the boundaries of the extraction region, and the blue crosshair within it represent the positions of the 1D slices displayed to the left. The bottom two plots contain the extracted data. The frequency domain data on the left displays the data prior to IFT. If symmetrization is not used, then it is identical to the F_1 slice in the top row, otherwise it displays the symmetrized peak. The time domain data on the right is the reconstructed FID calculated from inverse Fourier transformation of the plot to its left. All of these plots are updated in real time while the user defines the extraction region.

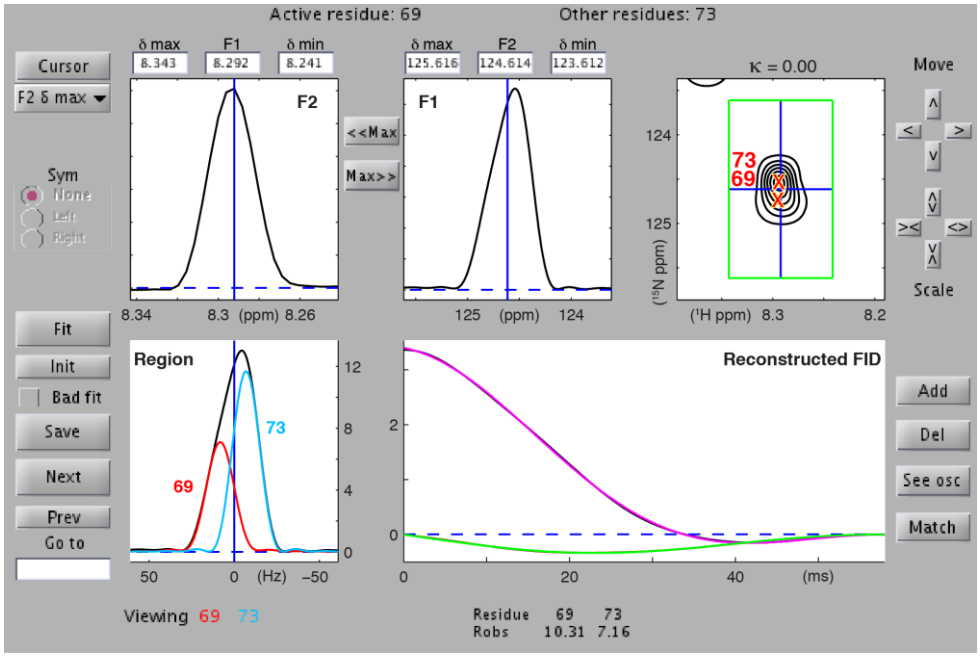

Figure 4.3 The FT/IFT dialog White boxes represent user-editable fields. Labels have been edited for clarity. Plots, from left to right and top to bottom, include: an F_2 slice from the spectrum, an F_1 slice from the spectrum, a contour plot of the 2D spectrum, the extracted region prior to inverse Fourier transformation and the reconstructed FID. The real (black) and imaginary (brown) parts of the reconstructed FID are both displayed. The green box surrounding the peak in the 2D contour plot represents the limits of the region of interest. The F_2 slice plotted at the top left corresponds to the horizontal line of the blue crosshair within the region-bounding box, and the F_1 slice at the top-middle corresponds to the vertical line. Users may "Move" and "Scale" the bounding box with the respective buttons to the right of the contour plot. Alternatively, users may specify the limits of the box and the positions of the slices within it using the editable fields above the F_1 and F_2 plots. Additionally, these fields can be defined interactively with a data cursor that is activated from the drop-down menu at the left. This menu specifies which field is being defined (here, "F2 δ max") Symmetrization may be applied using the "Sym" button group. The symmetrized peak is displayed in the "Region" plot at the bottom left. Alternatively, as has been done in this example, the user may add a second residue to the region using the "Add" button. Residues may be removed using the "Del" button. The active residue is listed at the top left, and any other residues added to the region are listed at the top right. Once satisfied with the reconstructed FID, the user presses the "Fit" button. The resulting fitted values of R_{obs} are displayed at the bottom right. The real (magenta) and imaginary (green) curves of the fit are plotted on top of the experimental reconstructed FID for comparison. The Fourier transform of the fitted curve is overlaid on the "Region" plot (not shown). Rather than plot the sum of all oscillators, individual oscillators may be Fourier transformed and overlaid on the "Region" plot (cyan & red peaks) using the "See osc" feature. The individual oscillators are displayed next to "Viewing". Oscillators are matched to their assigned residues using the "Match" dialog. The user may edit the initial fitting parameters using the "Init" button. If a satisfactory fit cannot be obtained, the residue may be marked as a "Bad fit" with the corresponding checkbox. Results must be saved ("Save") or discarded before the user may move to another residue using "Next," "Prev," or "Go to."

In order to maximize the SNR, the F_1 slice must be positioned at the height of the active peak in the F_2 dimension. Additionally, the user should be sure to capture the entire line-shape within the F_1 region. Truncating the base of the peak can lead to spurious oscillations in the reconstructed FID and systematic bias in the fitted rates (see Figure 4.5 and section 4.4.2). The boundaries of the region along F_2 only need to be large enough for unambiguous identification of the maximal signal intensity in F_2. The region may be moved, expanded or contracted in both dimensions of the contour plot using the controls to its right ("Move" and "Scale" in Figure 4.3).

In cases of partial overlap between peaks in F_1, special consideration must be taken when defining the region boundaries. The user may opt to accept some truncation of the active residue signal prior to IFT in order to separate the peaks; however, as mentioned above, this strategy can introduce systematic errors (see Figure 4.5 and section 4.4.2). Alternatively, the user can test if symmetrization resolves the overlap issue. SARA includes an option to reconstruct a full peak using only one half of the actual peak and can be accessed with the button group on the left side of the FT/IFT dialog ("Sym" in Figure 4.3). The symmetrization option may be beneficial in cases of slight overlap, i.e. when half of the signal is not perturbed by overlapping signals. However, symmetrization can introduce bias in rate measurements when the peak maximum does not coincide exactly with a point in the spectrum. Such a problem may be minimized with extensive zero-filling prior to Fourier transformation (see Figure 4.7 and section 4.4.4). To accomplish this SARA selects the single t_1 FID corresponding to the signal to be symmetrized,

increases the amount of zero-filling for this FID, and Fourier transforms this FID before applying symmetrization. Thus, symmetrization is not applied to signals taken from the low resolution, user-loaded spectrum but rather to signals benefiting from a much higher digital resolution. The procedure alleviates the need for extensive zero-filling of the entire dataset, thus saving disk space and accelerating calculations for other methods of analysis. The amount of zero-filling is estimated automatically to ensure that the bias introduced by symmetrization does not exceed 0.6% for rates of 5 s^{-1} and above. Clearly, this feature does not prevent bias originating from a partially overlapping peak that may contribute to the area of the signal used for symmetrization. In particular, it may be difficult to verify that half the line-shape of a signal is not perturbed by an overlapping signal when investigating two signals with a large difference in intensity. Users are advised to exercise the symmetrization option with care.

If all of these solutions fail, or if the overlap is too strong, the user may instead expand the region to contain both of the overlapping residues and fit two oscillators to the reconstructed FID instead of one. Pressing the “Add residue to region” button (“Add” in Figure 4.3) will search the peak list for residues positioned within the contour plot box and prompt the user to add them to the region. When fitting in the next step, SARA will determine the number of oscillators to use based on the number of residues identified in the region. Such an implementation of the FT/IFT procedure results in a reduction in the number of oscillators that are fitted when compared to the MP method. The major advantage is that the results of the

fit displayed by SARA are more simple to analyze and, hence, more likely to reveal poor fitting.

Step 2 – Optimizing data fit

The data is fit in the time-domain rather than the frequency domain. To transform the F_1 slice from the frequency domain to the time domain SARA extracts the region, performs symmetrization if requested, centers it at zero frequency, zero-fills the spectrum to the full spectral width, performs an inverse Fourier transform and truncates the resulting FID to the number of points acquired K (e.g. TD1, ni, etc.).

After pressing the “Fit” button, the reconstructed FID is fit to a sum of damped oscillators. The number of oscillators N included in the fit is determined based on the number of residues that were added to the region in the previous step. Each oscillator n is defined by an amplitude (A_n), a frequency (ω_n) and a decay rate ($R_{obs,n}$). By fitting the frequency of signals in the reconstructed FID, even when only one signal is present, we relax any requirement that signals be perfectly centered prior to IFT. The initial amplitude of each oscillator is set to the value of the reconstructed FID’s first point (i.e. at t = 0) divided by the number of oscillators N. The initial decay rates default to 5 s^{-1}. The initial frequencies are spread uniformly within the boundaries of the region, e.g. if there are two oscillators in a 300Hz region centered at zero then the initial frequencies will be set to +50 Hz and -50 Hz. As in the MP protocol, the fitted function is the sum of damped oscillators. However, it now contains a minimal number of oscillators, N, while being constrained to the same number of data points (K). In addition, the target

function is now multiplied by the apodization function, G(t), used for processing the spectrum:

$$F(t) = G(t) \sum_{n=1}^{N} A_n e^{(-R_{obs\,n} + i\omega_n)t} \tag{4.14}$$

The optimized parameter χ^2 is identical to (4.8 of the MP method. The resulting rate is a fit of R_{obs} in the current reconstructed FID only; the rate R_i is subsequently obtained by globally fitting both accordion spectra simultaneously (Step 3).

In cases where two or more oscillators are fit to the reconstructed FID, the sum of the oscillators is overlaid for comparison with the experimental data. Alternatively, the individual oscillators may be viewed using the "See individual oscillators" feature ("See osc" in Figure 4.3). The decay rates resulting from the fit are displayed for all residues present in the region. However, the fitting procedure is such that a set of fitted parameters (A, ω, R_{obs}) may not be assigned to the correct residue. The user can assign the correct oscillator to the correct residue using the "Match residues to oscillators" feature ("Match" in Figure 4.3). In cases of heavy overlap, the user should be sure to interpret the results with care, as the accuracy and precision of fitting decreases with increasing overlap.

At this stage, Steps 1 and 2 may be iterated to improve the fit. The quality of the fit is assessed by visual inspection of calculated and experimental points in the time and frequency domains (Figure 4.3). If the fit is not acceptable, the user may first try to adjust the initial estimates of the parameters to be fit. Users may also opt to fix a subset of the fitting parameters using values known through other,

independent measurements (e.g. peak frequency). Both of these strategies may be accessed using the initial parameter dialog ("Init" button in Figure 4.3). If the fit remains unsatisfactory, the user may return to Step 1 and adjust the parameters of the extraction region. If the result still does not meet expectations, the user may mark it as a "Bad fit" before moving on.

Step 3 – Global fit

Because the two accordion spectra have different values of κ, the width of signals along F_1, and consequently, the user-defined regions will be different for each spectrum. Therefore, for a given residue, the user must repeat Steps 1 and 2 on the second accordion spectrum prior to reaching step three. The third step preforms a global fit of both reconstructed FIDs for a given residue in order to extract R_i. Global fitting is initiated from the main SARA dialog. The global target function for each reconstructed FID is defined in much the same way as in the MP method, except that now it is multiplied by the apodization function G(t).

$$F_p(t) = G(t)\sum_{n=1}^{N} A_n e^{(-R_n^* + i\omega_n)t} e^{-\kappa_p R_{i\,n} t} \tag{4.15}$$

The optimized parameter is the same as in equation 4.13, except that now the experimental data I_p is a reconstructed FID rather than a slice in the acquired interferogram. For each oscillator, the average amplitude and frequency of the two test fits performed in step 2 are used as the initial values for the global optimization. The initial values for R_i and R^* are derived from the test fits of R_{obs} and are calculated using equations 4.9 and 4.10.

Step 4 – Monte Carlo error analysis

SARA offers a Monte Carlo error analysis dialog for the FT/IFT as well. As with the MP methods, the user must first establish the level of noise in each experiment and may specify the number of rounds of the Monte Carlo and the residues for which it will be performed. In each round of the Monte Carlo SARA adds traces of pseudo-random noise to the reconstructed FIDs and repeats Step 3. The resulting standard deviations of R^* and R_i are stored and added as error bars when plotting results.

4.4 Comparison of protocols

When determining the ideal fitting procedures for a given protein, the user should consider the advantages and disadvantages of each fitting method. In the MP method, the absence of spectral manipulation in the t_1 dimension prevents any loss of signal or introduction of artifacts, and therefore the MP method has a higher fitting precision than the FT/IFT method (see below for an example). Nevertheless, the MP method fits many signals simultaneously and signals overlapping in t_1 may produce erroneous results that go unnoticed. The FT/IFT method provides an opportunity for users to more closely inspect their data in order to identify inconsistencies and special circumstances. Once problems occurring with the MP method are identified with FT/IFT, the user may opt to re-analyze the data with the interactive MP method, albeit with input parameters estimated with FT/IFT. This section describes in detail the accuracy and precision of relaxation rates obtained

with the various protocols offered by SARA and describes how MP and FT/IFT can be used in concert to minimize bias while maximizing precision.

4.4.1 Accuracy of the MP protocols

One example from ubiquitin captures many of the features that distinguish the MP and FT/IFT fitting methods. An F_1 slice taken from the HN spectrum of our ubiquitin sample and centered on residue 49 (the active residue) also contains four other residues. In particular residue 49 is close to the strongly overlapped residues 31 and 72. When analyzing the corresponding t_1 FID with the MP method, the estimated number of oscillators provided to the HSVD algorithm was five. However, the HSVD algorithm yielded four oscillators to describe the data; only a single oscillator was predicted to represent the two strongly overlapped peaks. Subsequent non-linear fitting using four oscillators yielded results that appeared valid. However, concerns regarding the accuracy of rates in the presence of nearby, overlapping residues prompted further investigation. Indeed, in a simulated reconstruction of the data from these residues, the rate of the active residue 49, which itself is not subject to overlap, was unexpectedly biased by 3.4% when residues 31 and 72 were fit with a single oscillator (Figure 4.4 (a, b)). Accuracy for the rate of residue 49 was recovered, however, if the number of oscillators was increased to five and the overlapped signals were fit with two oscillators rather than one (Figure 4.4 (c, d)). Hence, the combination of two oscillators more completely accounts for the presence of the overlapped signals and prevents compensation by neighboring, isolated oscillators during the fit. Even so, the rates obtained for the strongly overlapped residues themselves are

extremely unreliable, as seen by a simple visual inspection of Figure 4.4 (d). While the occurrence of such overlapping residues can be predicted in a fully assigned protein, the blind application of the MP method to unidentified overlapping residues could lead to systematic bias in the fitted parameters. Indeed, the fits of both FIDs and their Fourier transforms appear deceptively good in Figure 4.4, owing to the presence of other, well-fit residues, and such erroneous results may well go unnoticed. In fact, visual inspection of individual oscillators might suggest that using four oscillators provides superior results Figure 4.4 (b) vs. (d)). Clearly, when analyzing experimental data such a situation would likely go unnoticed; four oscillators would be used in the fit, and the rate of residue 49 would be inaccurate. Fortunately, the FT/IFT method provides an alternative means of analysis that reveals inconsistencies and may often provide relief, as discussed in the following section.

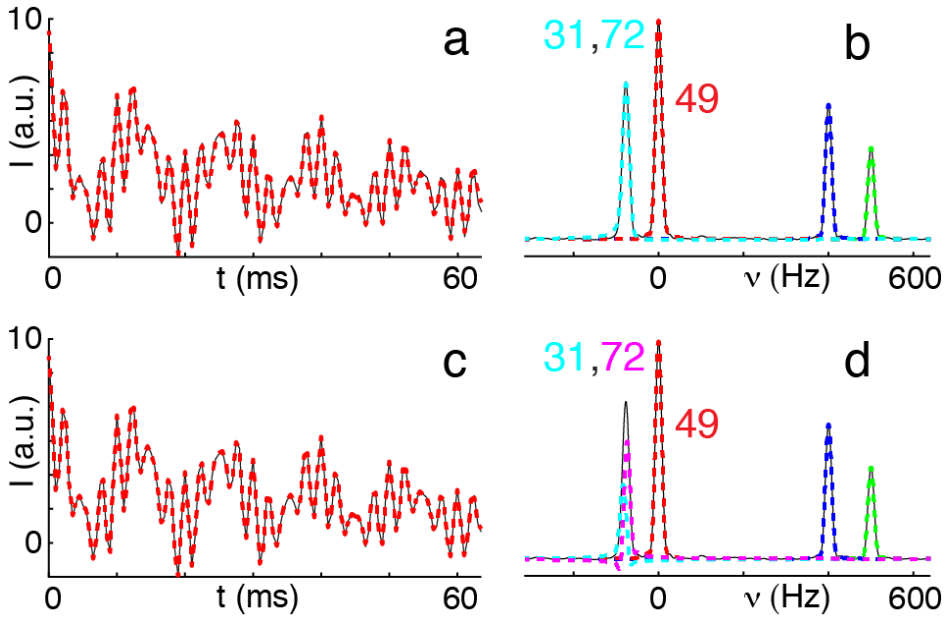

Figure 4.4 Inaccuracies of the automated MP method The four panels contain simulated data (solid) overlaid with their corresponding fits (dashed). The simulated data reproduces the t_1 FID and spectrum slice of residue 49 in ubiquitin. It contains five damped oscillators corresponding to residues 4, 18, 31, 49 and 72. The frequencies of 31 and 72 are nearly degenerate. a) Time domain data (black) overlaid with a fit including four damped oscillators (red). b) Fourier transform of the FID in (a) (black) overlaid with the Fourier transform of each individual damped oscillator (colors). Visual inspection does not indicate inaccuracy for residue 49. c) Same time domain data as in (a) (black) overlaid with a fit (red) using *five* damped oscillators. d) Fourier transform of the data in (c) (black) overlaid with the Fourier transform of each individual oscillator of the fit (colors). Residues 31 and 72 seem to give poorer results. While visual inspection of panels (a-d) for residue 49 indicates similar performance between fitting with four and five oscillators, the fit in panels (a) and (b) is inaccurate by 3.4%. This figure illustrates the importance of close inspection by the user when applying non-linear fitting methods. The FT/IFT method can be used to closely inspect and evaluate multiple fitting approaches in these types of situations. The fitting strategy and parameters can then be implemented in the interactive MP method for maximum fitting precision.

4.4.2 Accuracy of the FT/IFT protocol

The FT/IFT method provides an alternative that both isolates the fitting of signals from one another and allows closer inspection by the user, facilitating the detection of poor fits. The procedure may allow users to overcome the limitations of the MP method. When analyzing the example discussed above with the FT/IFT method, residue 49 was extracted together with residues 31 and 72 and fit using

only two oscillators, corresponding to the default number of four oscillators used in the MP method. Comparison between the rates obtained by MP and FT/IFT revealed a discrepancy of 0.6 s^{-1} between the two, with the rate obtained by FT/IFT biased by only 1.1 % whereas that of MP suffered from 3.4% bias. This improvement in accuracy further demonstrates that the error in the rate estimated by MP originated from the cumulative effect of numerous and overlapping oscillators in the same t1 FID; when only two oscillators needed to be fit rather than four, the rate of 49 became more accurate. A markedly better fit could be observed with the FT/IFT method when using three oscillators rather than two, providing a rate accurate within 0.03 %. The simplified interferogram produced by the FT/IFT method simultaneously increases the chances of identifying such an issue and provides the correct solution. Alternatively, one could isolate residue 49 using the symmetrization feature, taking advantage of the minimal overlap between 49 and 31/72. With symmetrization, the rate was accurate within 1%, demonstrating that the rate of the symmetrized signal was closer to that obtained with three oscillators than that obtained with two oscillators using standard FT/IFT processing. Although FT/IFT, with or without symmetrization, provides a more accurate answer in this example, this trait cannot be generalized, and we seek only to highlight that FT/IFT can be used as a tool to identify discrepancies and help resolve them. Indeed, when the parameters obtained by FT/IFT were supplied to the interactive MP procedure, the estimated rate became accurate to the fourth decimal place and benefited from the increased precision of the MP method. We note that the extraction and symmetrization procedures themselves can introduce

bias, and this section describes the sources of these systematic errors and strategies to mitigate them.

When determining the spectral region boundaries prior to extraction, users should be sure to capture the entire line-shape within the region. Figure 4.5 displays the normalized bias of a fitted rate as a function of the relative region width extracted. To determine these quantities, simulated FIDs were created with various signal-to-noise ratios and relaxation rates. A cosine-squared window function was applied, and the FT/IFT procedure was performed for extracted regions of various widths. The entire procedure was then repeated in a 1000 round Monte Carlo simulation. The absolute bias δ is calculated as the difference between the mean of the measured rates (R_m) determine by the FT/IFT method and the true rate used to simulate the signal (R_t), $\delta = R_m - R_t$. The normalized bias is calculated as the absolute bias δ divided by the simulated rate R_t. The relative region width is defined as the extracted region width divided by the full-width at half-max (FWHM) of the apodized and Fourier transformed peak. These simulations reveal that extraction regions which truncate the line-shape lead to systematic underestimation of the relaxation rate. As expected, the bias is dramatic when the extracted region approaches the FWHM and reduces to zero as the extraction region grows.

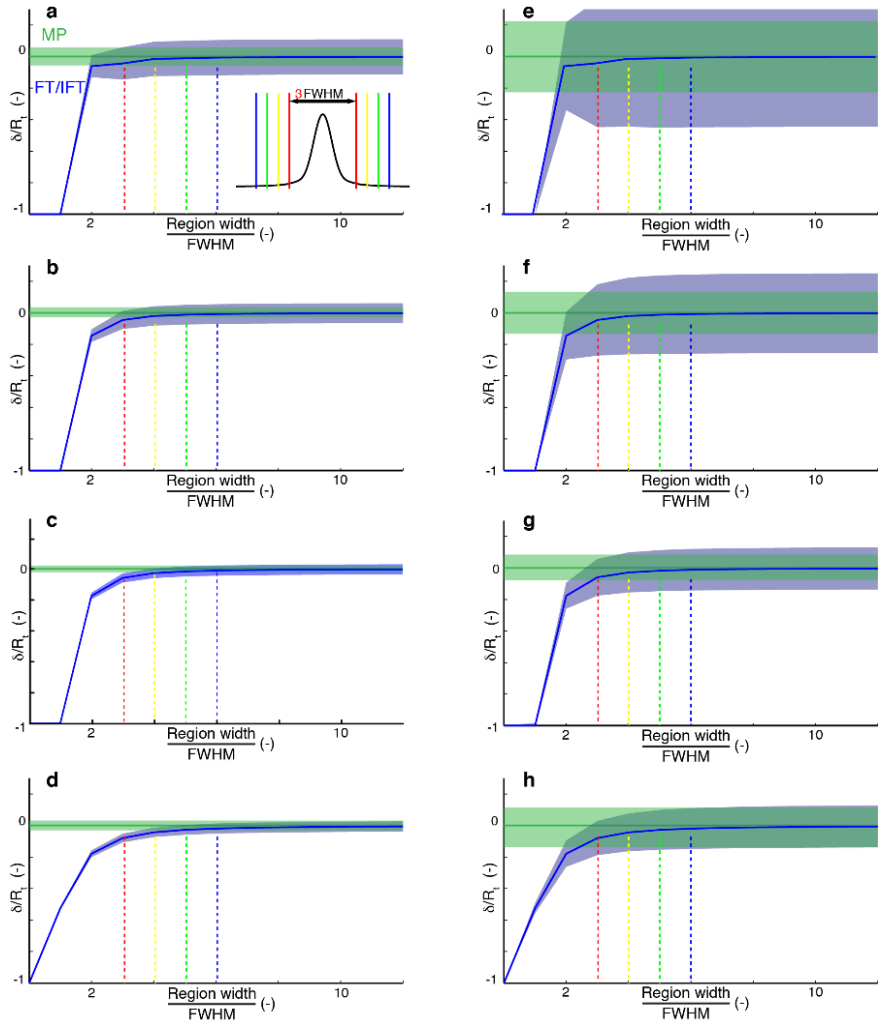

Figure 4.5 Normalized bias and relative region width with FT/IFT. Normalized bias, δ/R_t, was calculated as the absolute bias $\delta = R_m - R_t$ divided by the true rate used to generate the signal, R_t. R_m is the mean of the fitted rates, R_{obs}, over a Monte Carlo simulation. Normalized error was calculated as the standard deviation of the Monte Carlo simulation, σ, divided by the true rate, R_t. Bold centerlines denote the normalized bias while the shaded regions represent the extent of the normalized error (±σ). The bias and error of the FT/IFT method (blue) is compared with that of the MP method (green), which is displayed identically at all window widths for visualization. Fits of R_{obs} were calculated over a 1000 round Monte Carlo performed on a decaying exponential defined by 128 complex points at intervals of 500μs with signal-to-noise ratio of 20 (a-d) or 5 (e-h), and with decay rates of 5 s^{-1} (a, e), 10 s^{-1} (b, f), 30 s^{-1} (c, g) or 100 s^{-1} (d, h). During each round of the Monte Carlo, pseudo-random noise was added to the pure signal and the resulting FID was fit using the MP method. The FID was then apodized with a cosine-squared window function, zero-filled to 4096 points, Fourier transformed, and fit using the FT/IFT method at various window widths. Relative region widths were chosen as integer multiples of the full width at half max (FWHM) as measured on the noiseless peak. The inlay displays the corresponding peak line-shape. Solid bars in red, yellow, green and blue indicate region widths of 3, 4, 5 and 6 times the FWHM respectively. Dashed lines in matching colors indicate the respective inaccuracy at each level of truncation.

Inspection of Figure 4.5 also reveals that the error obtained by the Monte Carlo analysis does not increase as the peak is truncated. Consequently, rates extracted from substantially truncated signals may appear deceptively precise while being strongly biased. That is, the error-bars associated with such a rate would not account for the systematic underestimation induced by truncation, and the probability that the true rate would be contained within the error-bar-interval would be reduced.

More generally, when assessing the influence of bias on the reliability of relaxation rate measurements, it is important to consider the any bias relative to the measurement's error. To quantify this effect, we introduce a parameter δ/σ that reports on both bias (δ) and precision as measured by the standard deviation (σ) of fitted rates in a 1000 round Monte Carlo analysis. This parameter provides insights on the confidence level associated with the error-bar-interval of a fitted rate. To understand this parameter we consider the following. For a given distribution of fitted rates, bias moves the mean of the distribution away from the true rate. A value of $\delta/\sigma = 0.5$ implies that the mean of the distribution deviates from the true rate by half of the distribution's standard deviation. Thus, the probability of finding the *true* rate within the fitted rate's error bars is reduced. For example, if a fitted rate were drawn from a Gaussian distribution with zero bias and standard deviation σ, error bars of $\pm\sigma$ about the fitted rate would describe a 68% confidence interval. A bias of $\delta = 0.5\sigma$ ($\delta/\sigma = 0.5$) reduces the confidence level of the same error-bar-interval to 62%. Likewise, a value of $\delta/\sigma = 1$ further reduces the confidence level to 48%.

Figure 4.6 displays the ratio of the bias (δ) to the standard deviation (σ) as a function of SNR for various levels of truncation and at different relaxation rates. Here, SNR is defined as the amplitude of the simulated signal in the time domain divided by the standard deviation of the time domain noise. The horizontal dashed line highlights a value of $\delta/\sigma = 0.5$, which we consider to be an acceptable level of bias. At low SNR, the experimental error dominates and any bias introduced from truncation has little effect. On the other hand, at high SNR increasing truncation has a substantial effect. In these cases, as much of the signal as possible must be extracted to ensure that systematic error does not corrupt the relaxation rates.

Furthermore, as we increase relaxation rate, the concomitant increase in the slope of the lines indicates that larger rates are more prone to inaccuracies caused by truncation. However, this analysis compares signals of different relaxation rate at equal signal-to-noise ratios. In practice, signals with higher relaxation rates have reduced SNRs, tending to compensate for this trend.

In summary, the FT/IFT procedure may introduce bias for excessive truncation of signal line-shapes, and this bias may not be accounted for by the experimental error associated with the estimated rate when truncation is performed at high SNR. In such cases, to restore the rate's validity, users should either increase the width of the extracted region (and consider including additional oscillators if necessary) or they should use the symmetrization feature if possible. Figure 4.6 provides an empirical means to evaluate the validity of the estimated rates using FT/IFT.

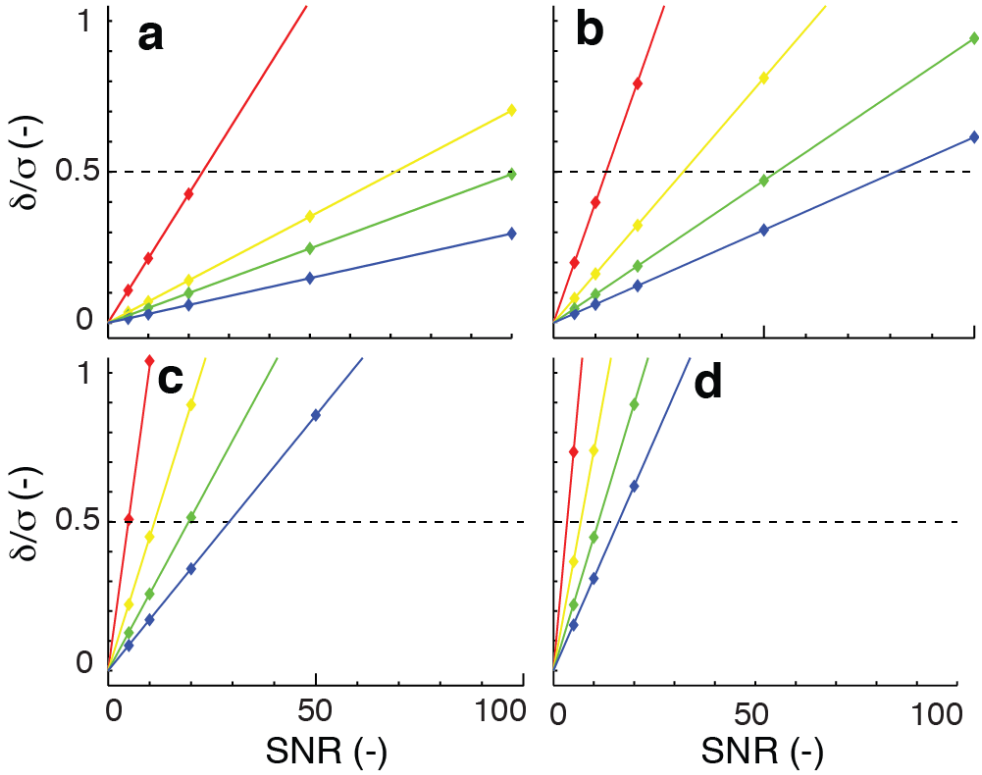

Figure 4.6 The variation of bias and error with FT/IFT. Data is taken from simulated signals with decay rates of (a) 5 s^{-1}, (b) 10 s^{-1}, (a) 30 s^{-1} and (b) 100 s^{-1}. In each plot, the fixed value of bias (δ), which is independent of SNR, is divided by the standard deviation (σ) of 1000 fits of the relaxation rate at different values of SNR. The parameters for each round of the Monte Carlo simulation are as described in Figure 4.5. The colored lines indicate region widths of 3 (red), 4 (yellow), 5 (green) and 6 (blue) times the FWHM. The dashed line indicates a bias that is equal to half the standard deviation (δ/σ = 0.5). If these rates were drawn from a Gaussian distribution, this value would correspond to a reduction in confidence of the error-bar-interval (±σ) from 68% to 62%. For all rates, increasing the extracted region width increases the confidence level by reducing bias. At low SNR, the inherently large error dominates and even substantial truncation has little effect on the level of confidence of the error-bar-interval. The validity of rates obtained with FT/IFT is most questionable for signals with large relaxation rates yet high SNR, where the δ/σ ratio would be large, reflecting that the error-bar-interval would be unlikely to contain the true value. In practice, such a situation is unlikely to occur because fast relaxation leads to reduced SNR.

4.4.3 Precision of the FT/IFT protocol

We have already discussed how the noise influences the precision of rates obtained by FT/IFT. However, the precision of these rates is also affected by a reduction of the signal intensity stemming from apodization. The ability to isolate

(groups of) signals in the frequency domain requires apodization of the time domain data. Indeed, the t_1 FID must be free of truncation to prevent "sinc-wiggle" artifacts that would otherwise make the extraction procedure inapplicable. This constraint is also present for other methods that extract relaxation rates following Fourier transform in t_1, such as line-shape analysis[17] or simply line-width measurements[15]. In these previous works the authors restricted their analyses to data acquired either with large values of κ[15] or subject to rates that were intrinsically large[17], leading in both cases to fully relaxed FIDs that required no apodization. Unfortunately, this solution is not applicable in general because most proteins display a large dynamic range of relaxation rates. To prevent sinc-wiggles, the user would have to sample relaxation times dictated by the smallest rates. However, such a constraint would greatly reduce the signal-to-noise ratio of signals subject to larger relaxation rates, for which the majority of the data acquired would be noise. In a typical, well-folded protein, slowly relaxing residues are a minority, located in loops and terminal regions of the polypeptidic chain, and the experiment would be sub-optimal for a majority of signals. Therefore, it is preferable to design the acquisition to maximize the sensitivity for all residues and to overcome truncation artifacts with apodization. Indeed, Lefèvre and co-workers have already implemented a procedure that includes apodization, allowing accordion data to be analyzed in the frequency domain even with a value of $\kappa = 0$ (i.e. a reference experiment). However, their solution still required that all signals in an F_1 slice be fit simultaneously and did not take advantage of the frequency domain separation achieved by the Fourier transform.

To account for the apodization, it may appear efficient to divide the reconstructed FID by the apodization function prior to fitting, thus restoring the signal-to-noise ratio and providing access to the simpler target functions of the MP procedure (equations 4.7 & 4.12). However, such a process frequently introduces singularities in the reconstructed FID due to division by numbers tending toward zero. In contrast, inclusion of the apodization function in the target function requires only a point by point multiplication and does not increase the number of fitted parameters. Unfortunately, we note that apodization results in a loss of signal intensity that translates into a decrease in the precision of fitted parameters. For example, a signal with a relaxation rate of 5 s^{-1} is accompanied by an error of 0.056 s^{-1} with our implementation of the MP method. The error increases to 0.11s^{-1} when fitting with the FT/IFT method, corresponding to a two-fold loss in precision. Nevertheless, apodization is necessary for a broad application of our FT/IFT method.

4.4.4 Accuracy of FT/IFT symmetrization

The symmetrization feature within the FT/IFT protocol provides two potential advantages. Given partially overlapped peaks, it allows users to make a comparison between fitting both peaks simultaneously and fitting each signal separately after symmetry reconstruction. Symmetrization may also help users identify overlapped residues that had otherwise gone unnoticed. In such a case, a seemingly isolated signal would give rise to two substantially different rates when reconstructing from each half of the peak. The identification of overlapped peaks is clearly valuable in the context of protein functional studies, as improper fitting

can result in relaxation rate outliers and subsequent erroneous mechanistic interpretations. However, symmetrization can introduce additional biases, and SARA implements a special procedure to minimize them.

Symmetrization produces erroneous results when the peak maximum does not coincide exactly with a point in the spectrum, but this problem can be mitigated with extensive zero-filling prior to Fourier transformation in t_1. The most severe bias is introduced when a peak's maximum is centered exactly between two points in the spectrum. The magnitude of this inaccuracy is then a function of the distance between the true center of the peak and the closest point in the spectrum. That is, biases introduced by symmetrization will depend on the digital resolution of the spectrum. To quantify this effect and to design a solution in SARA, we have performed simulations at various relaxation rates and signal-to-noise ratios for signals subject to this worst-case scenario.

Figure 4.7 displays the accuracy of rates obtained by symmetrization as a function of the number of points encompassed by the FWHM. For each peak, symmetrization was performed by using either the right half, which contains one more point (Figure 4.7 (b) top), or the left half, which contains one fewer point (Figure 4.7 (b) bottom). The former strategy artificially broadens the signal and overestimates the rate Figure 4.7 (a), green) while the latter narrows the signal and underestimates the rate (Figure 4.7 (a), blue). Comparison of simulations for rates of 1 s^{-1}, 5 s^{-1}, and 100 s^{-1} (Figure 4.7 (a)) reveals that larger relaxation rates are the least affected by variation in spectral resolution, whereas smaller rates require a large digital resolution to be accurate. This observation simply reflects

that a sharp signal will be more sensitive to inaccuracies stemming from a symmetrization not performed around its true maximum. Figure 4.7 (a) reveals that fitting after symmetrization remains accurate to within 2.75% for a rate as low as 1 s^{-1} if the signal's FWHM contains 800 points (vertical red dashed line). This criterion was hence chosen as a means to minimize inaccuracies for all rates. Figure 4.7 (c) shows normalized bias (δ/R_t) as a function of relaxation rate under the 800-points-per-FWHM condition. All rates greater than or equal to 5 s^{-1} remain accurate within 0.6% following symmetrization.

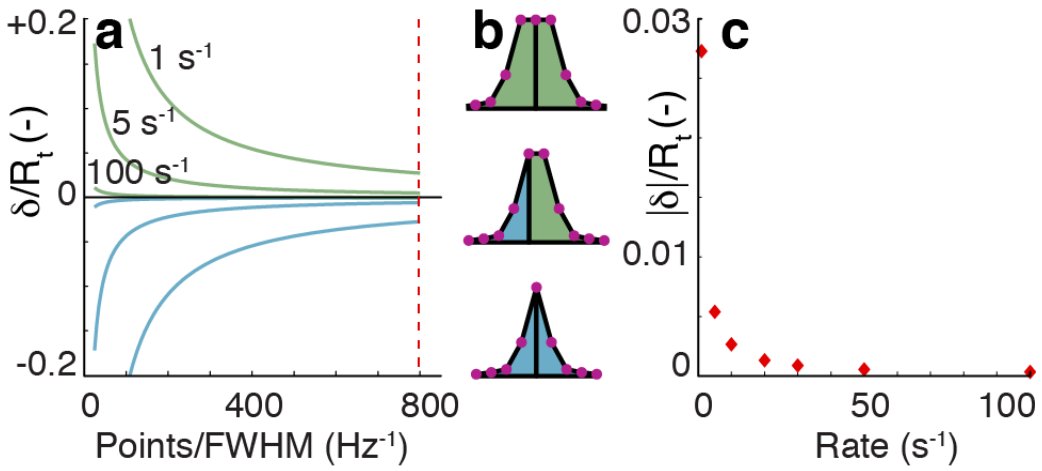

Figure 4.7 Bias introduced by symmetrization. The worst-case scenario for symmetrization occurs when the true peak maximum lies exactly between two points in the spectrum (purple dots, (b) center). In this case symmetrization would reconstruct a peak that is too narrow ((b) bottom, blue) or too wide ((b) top, green). In either case, the bias decreases when increasing the extent of zero-filling prior to Fourier transformation. Panel (a) displays the bias when fitting either the narrow (blue, bottom) or wide (top, green) symmetrized signal as a function of the number of points spanning the FWHM of the peak. Symmetrization simulations were performed using the same parameters as in Figure 4.5 and Figure 4.6 except that in each case the peak frequency was offset from zero by one half of the digital resolution. The simulated decay curves were apodized, zero-filled to multiple values based on the known FWHM and subsequently Fourier transformed, symmetrized, inverse Fourier transformed and fit. Curves are displayed for relaxation rates of 1, 5 and 100 s^{-1}. The dashed line in panel (a) indicates the amount of zero-filling used in panel (c). Panel (c) displays the magnitude of the bias as a function of the relaxation rate at a zero-filling level leading to 800 points per FWHM for each rate. Rates as low as 5 s^{-1} are accurate to within 0.6% under such conditions. SARA includes an automated procedure using this level of zero-filling to minimize the bias introduced by symmetrization.

In a manner similar to that described for Figure 4.6, the validity of rates after symmetrization was assessed by monitoring the bias relative to the error as a function of SNR for various values of zero-filling (Figure 4.8). The plots reveal that, when symmetrizing signals with 800 points per FWHM, nearly all conditions result in an acceptable amount of bias ($\delta/\sigma > 0.5$). At this level of zero filling, signals with rates as low as 5 s^{-1} and SNR as high as 300 still generate errors-bar-intervals providing a high confidence level (above 62% when assuming a Gaussian distribution). Consequently, this zero-filling criterion was used to design a procedure to limit the bias introduced during symmetrization in an automated manner.

For any spectral width and apodization function provided by the user, SARA automatically calculates the amount of zero-filling needed for accurate symmetrization. More specifically, a signal decaying with a rate of 1 s^{-1} is simulated and apodized. It is then Fourier transformed so that its FWHM can be measured, and SARA calculates the amount of zero-filling needed to reach 800 points within this width (ZF_{sym}). After selecting one of the “Sym” options (Figure 4.3), SARA extracts the corresponding t_1 FID from the interferogram, zero-fills it to ZF_{sym} points, and re-Fourier transforms the data. Symmetrization is then performed on this single, high-resolution spectral trace before inverse Fourier transformation and fitting. This procedure ensures that symmetrization does not introduce inaccuracies higher than 0.6% for rates as low as 5 s^{-1}.

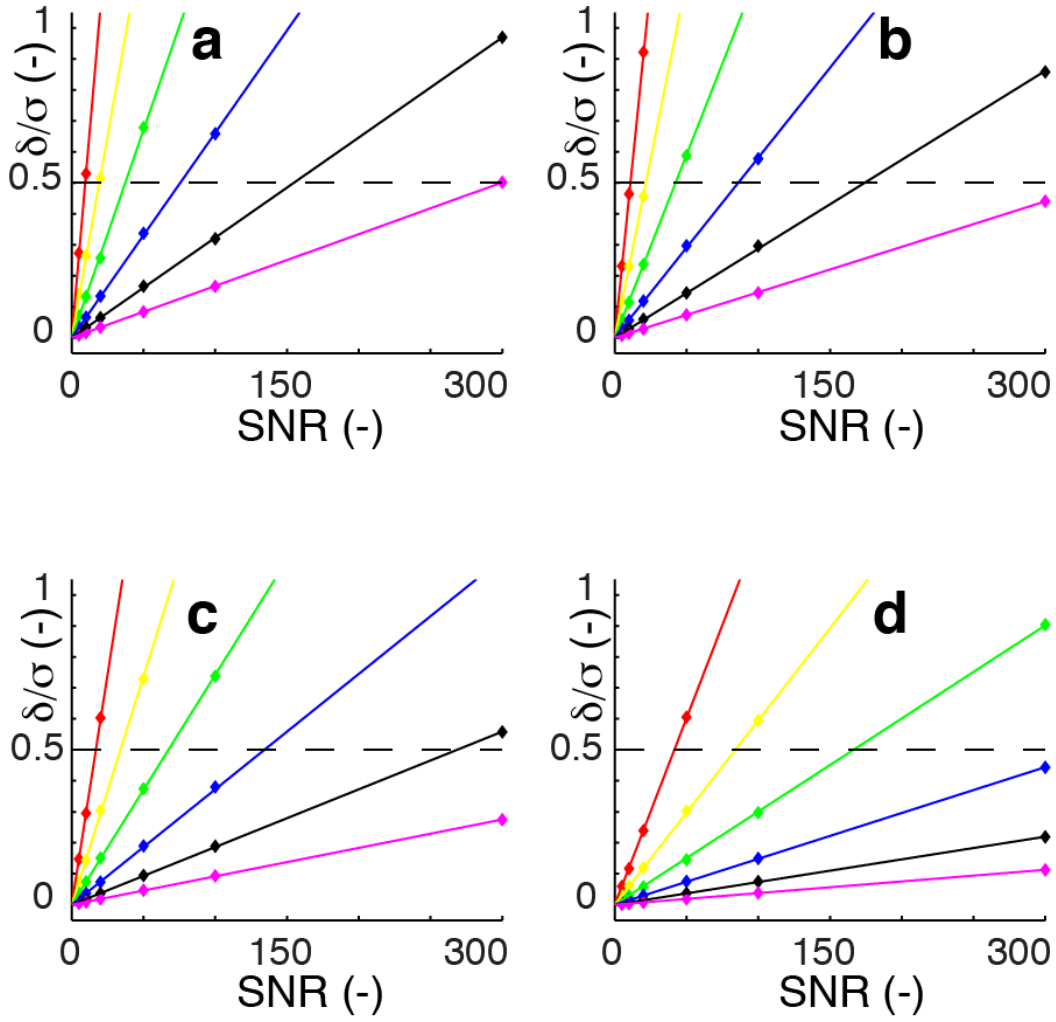

Figure 4.8 The variation of bias and error when symmetrizing Data is taken from simulated signals with decay rates of (a) 5 s^{-1}, (b) 10 s^{-1}, (c) 30 s^{-1} and (d) 100 s^{-1}. In each plot, the fixed value of bias (δ), which is independent of the noise amplitude, is divided by the standard deviation (σ) of 1000 fits of the relaxation rate at different values of SNR. The parameters for each round of the Monte Carlo simulation are as described in Figure 4.5. The colored lines indicate zero-filling levels which lead to 25 (red), 50 (yellow), 100 (green), 200 (blue), 400 (black), and 800 (magenta) points spanning the FWHM for a given rate. The dashed line indicates a bias that is half the standard deviation (δ/σ = 0.5). If these rates were drawn from a Gaussian distribution, this value would correspond to a reduction in confidence of the error-bar-interval (±σ) from 68% to 62%. For all rates, increasing the extent of zero-filling increases the confidence level by reducing bias (e.g. red (25 points per FWHM) vs. magenta (800 points per FWHM) lines). Rates as low as 5 s^{-1} remain valid at SNRs as high as 300 when zero-filling is applied such that 800 points span the FWHM of the peak (magenta line in (a)). Similarly, measurements of a rate of 1 s^{-1} remain valid at SNRs greater than 250 using the same 800-points-per-FWHM guideline (data not shown). SARA implements an automated procedure based on this criterion to ensure accurate results when applying symmetrization.

4.4.5 Complementary approach

Although the features offered by symmetrization and the FT/IFT method in general do come with caveats, they provide a necessary alternative to the MP method that can be used to validate rates and identify setbacks. In order to capture the advantages of both procedures, fitting strategies may be developed and tested using the FT/IFT method and subsequently implemented in the MP method to maximize precision.

In the end, we have found that the MP and FT/IFT procedures are highly complementary. The MP procedure has the indisputable advantage of analyzing the raw data and therefore maximizes precision in the fitted parameters because no apodization is necessary. However, the FT/IFT procedure may identify signals that were fit poorly in the MP protocol. In particular, the FT/IFT method overcomes limitations due to the cumulative effects of numerous and overlapping signals in the indirect dimension. Thus, we recommend beginning the analysis with the FT/IFT procedure, because it calls for the closest inspection by the user. This allows the user to clearly identify overlapping residues as well as “bad” signals (e.g. signals close to t_1 noise or axial peak artifacts). We then suggest using the user-interactive MP protocol to maximize precision. Users may minimize the risk of converging to a local minimum in the MP method by using the amplitudes, frequencies and rates obtained in the FT/IFT procedure as input to the optimization. Additionally, parameters known from independent measurements can be incorporated or held fixed during fitting. The local minima we encountered when applying the MP method to ubiquitin highlight the importance of user

intervention when applying non-linear fitting methods and emphasize the need for an interactive and user-friendly software environment such as SARA. Because SARA provides a visual comparison between the data and its fitted function in both procedures, any discrepancy between them can be identified and possibly resolved. We believe such a protocol allows researchers to analyze accordion data reliably and to clearly identify potential inaccuracies.

5 Molecular cross-talk between an NRPS carrier protein and its linkers

Modified portions of this text have been submitted for publication in *Angewandte Chemie International Edition*[90]

5.1 Structural interactions between $PCP1_{ybt}$ and its linkers

Biological systems often rely on protein domains separated by linker regions to partition or multiply functionality. For example, nonribosomal peptide synthetases (NRPSs) are microbial enzymatic systems that employ a modular, multi-domain architecture to incorporate substrates into secondary metabolites[22]. This assembly-line strategy, however, does not operate through a rigid molecular assembly but instead relies on structural fluctuations both within and between individual domains connected by linkers[27,91]. Importantly, carrier protein domains covalently tether substrates through a phosphopantetheinyl (PP) arm introduced on a conserved serine, and they were found to shuttle substrates between catalytic domains during synthesis[92,93]. Our lab recently provided evidence of molecular cross-talk between a carrier protein and its tethered substrate[94]. Here, we show that similar communication exists between the core of a carrier protein and its inter-domain linker regions. Using NMR we show that residues in the N-terminal linker interact with the domain core and modulate its dynamics. Further, by monitoring the protein's invisible unfolded state, we show that this linker region stabilizes the carrier protein fold.

5.1.1 Structure of $PCP1_{ybt}$

We determined the solution structure (PDB 5U3H) and fast dynamics (ps-ns) of $PCP1_{ybt}$, a peptidyl carrier protein (PCP) from Yersiniabactin Synthetase[33], including flanking residues linking $PCP1_{ybt}$ to its upstream adenylation and downstream cyclization domains (Figure 5.1). The NMR structural statistics for the bundle are summarized in Table 5.1. The core displays a four helix bundle commonly observed in carrier proteins[95], with the active serine S1439 at the N-terminus of α2 (* in all figures). Model Free analysis of ^{15}N NMR relaxation[65,96,97] provided us with order parameters S^2 to characterize ps-ns fluctuations of amide bonds. S^2 range from 0 to 1, and are sensitive to both the amplitudes and orientations of motional processes. A value of 0 indicates an unrestricted motion and 1 denotes either a motion coaxial with the amide bond or rigidity.

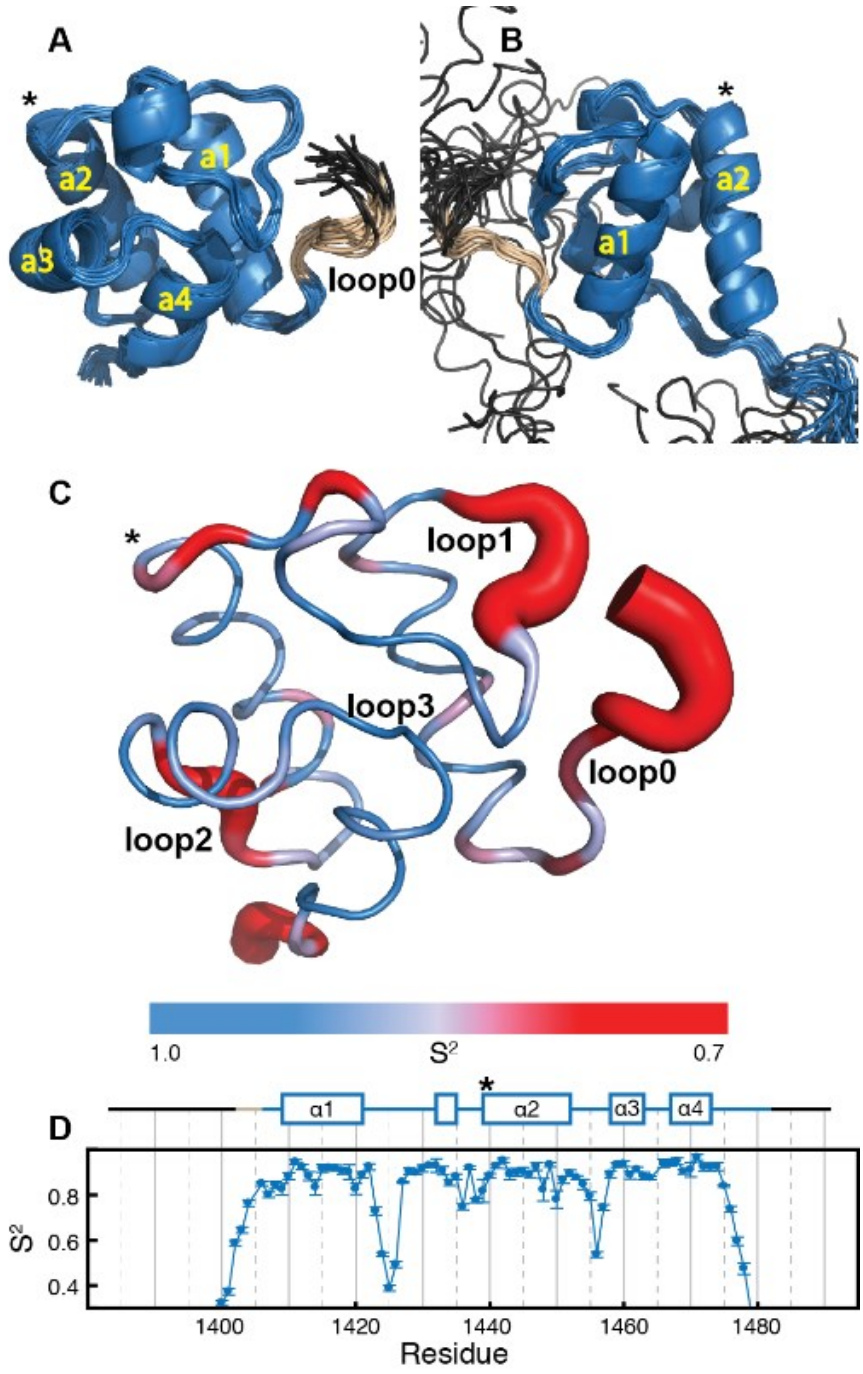

Figure 5.1 Structure and dynamics of $PCP1_{ybt}$ (1381-1491) (A) NMR bundle displaying residues 1400 to 1478. (B) Alternative view, also showing the disordered N- and C-termini residues in linkers. (C) 3D representation of $PCP1_{ybt}$ dynamics. A thicker width indicates a lower order parameter and increased ps-ns dynamics. The orientation and range of residues are as in (A). (D) Order parameters used in (C) and secondary structure. Colors in (A), (C), (D) are consistent and help define the constructs used in this work. Blue: 1406-1482 (protein core and emerging linkers), beige: 1402-1405 (contact region of loop0) and black: disordered regions in linkers.

NMR structure statistics for $PCP1_{ybt}$	
Violations (mean and s.d.)[a]	
Distance constraints (Å)	0.27 +/- 0.04
Dihedral angle constraints (º)	1.69 +/- 0.17
Max. dihedral angle violation (º)	2.09
Max. distance constraint violation (Å)	0.37
R. m. s. deviations geometry [b]	
Bond lengths (Å)	0.019
Bond angles (º)	1.2
Average pairwise r.m.s.d. [residues 1403–1477] (Å) [a]	
Heavy	0.75 +/- 0.07
Backbone	0.22 +/- 0.06
Ramachandran Statistics [c]	
Most favored	92.7
Additionally allowed	7.3
Generously allowed	0.0
Disallowed	0.0

Table 5.1 NMR structure statistics for $PCP1_{ybt}$. (a) from CYANA 2.1[54,55], (b) From PSVS[57], (c) from ProCheck[58]. The structure can be accessed with PDB code 5U3H.

The pattern of flexible residues observed in $PCP1_{ybt}$ (Figure 5.1 (C, D)) is reminiscent of those reported in aryl and acyl carrier proteins[94]. Loop2 and the N-terminus of loop1 are malleable, while the four helices are relatively rigid. Of particular interest, the residues preceding the active serine (1436-1438) display alternatingly low and high values of S^2, with the S^2 of G1437 10% and 15% higher than those of its predecessor and successor, respectively (panel (D)). This behavior may reflect the sensitivity of order parameters to anisotropic motions rather than an alternation of rigid and flexible residues. Acyl and aryl carrier proteins also display dynamics around the PP site, and malleability here is likely needed during the many molecular events involving these residues, including post-translational modifications[94]. Our results indicate that peptidyl carrier proteins also possess malleability in this region.

Although most of the N- and C-terminal linker regions of $PCP1_{ybt}$ are strongly disordered (Figure 5.1 (A) and Figure 5.11), there are three exceptions. First, the five, C-terminal linker residues following α4 interact with the core along loop2. Second, a short region approximately 20 residues upstream of α1 displays helical character, as evidenced by chemical shift indexing and weak nuclear Overhauser effect (nOe) cross-peaks. A recent study assigned these residues to the preceding adenylation domain[98], where they are expected to form a helix, and our results confirm this prediction. This region does not interact with the PCP core; its low S^2 indicates independent molecular tumbling (see Figure 5.11), and no nOe cross-peaks to the core are observed. In contrast, a third region immediately upstream of α1 displays extensive nOe's to the folded core, and its fast time-scale dynamics resemble those seen in loops rather than those seen in disordered linkers (Figure 5.1 (D)). Consequently, we refer to this region as loop0. Contacts between loop0 and the core occur principally through the proline residue P1405. Notably, proline is the most common amino acid at this position[98], and other PCP structures substantiate that observation. Out of 12 different Type I PCPs selected from the PDB, 11 contain a proline residue at this position and 9 of these display contacts between the proline and the core (Figure 5.2).

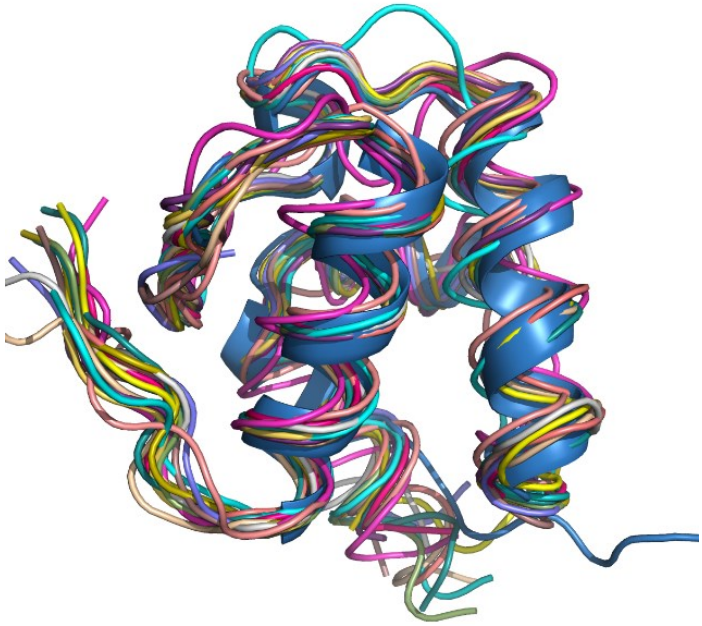

Figure 5.2 Loop0 contacts in other systems. $PCP1_{ybt}$ aligned with 16 other PCP structures from 9 different systems. Each structure is aligned to the first structure from the $PCP1_{ybt}$ bundle, shown in ribbon, and displays similar contacts between the protein's N-terminal linker and its core maintained by a proline residue. The following structures were included: 1DNY; 2GDW; 2JGP; 2ROQ; 2VSQ; 3TEJ; 2MD9 & 4MRT; 4PWV & 4PXH; 4R0M; 4ZXH, 4ZXI & 5T3D; and 5ES8 & 5ES9.

In $PCP1_{ybt}$, loop0 makes contacts to α1, loop3, and loop1. Loop1 plays an important role in communication with partner catalytic domains[95], prompting us to further characterize the influence of loop0 on the structure and dynamics of the folded core. Thus, we developed a second construct (blue in secondary structure diagrams throughout this chapter) that removes both N- and C-terminal linkers and excludes P1405, breaking the interaction between loop0 and the core.

5.1.2 Structural effects of removing loop0

To discern the influence of loop0 on the structure of the folded core, chemical shift perturbations (CSPs) were calculated between full length $PCP1_{ybt}$ (residues 1383 -1491) and truncated $PCP1_{ybt}$ (residues 1406-1482) for all

backbone and C^{β} nuclei (H^{N}, N, H^{α}, C^{α}, C^{β} and C'). The individual chemical shift values $\delta\nu$ in Hz are plotted in Figure 5.3 for each residue of the truncated construct. The single CSP $\Delta\Omega$ representing each residue in Figure 5.4 was calculated as the RMSD of the individual CSPs $\delta\nu$ [99,100].

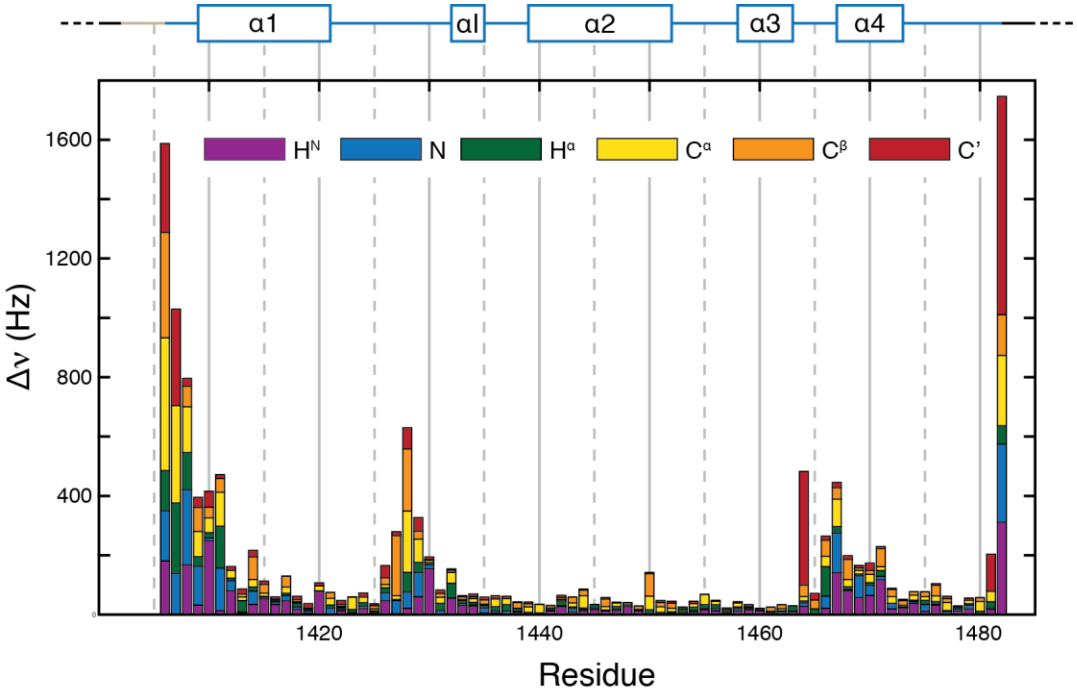

Figure 5.3 Stacked bar plot of individual CSPs between full length and truncated $PCP1_{ybt}$ for all backbone and C^{β} nuclei. The secondary structure of $PCP1_{ybt}$ is shown above the plot.

The CSPs localize primarily to the sites of truncation and to regions that contact loop0 (Figure 5.4), indicating that the structure of the core is largely unaltered. Only residue H1464, at the N-terminus of loop3, stands apart since it is distant from the contact point between loop3 and loop0. Notably, only the chemical shift of the carbonyl carbon is strongly affected (Figure 5.3), which may indicate (transient) hydrogen bond formation upon subtle rearrangements of loop3.

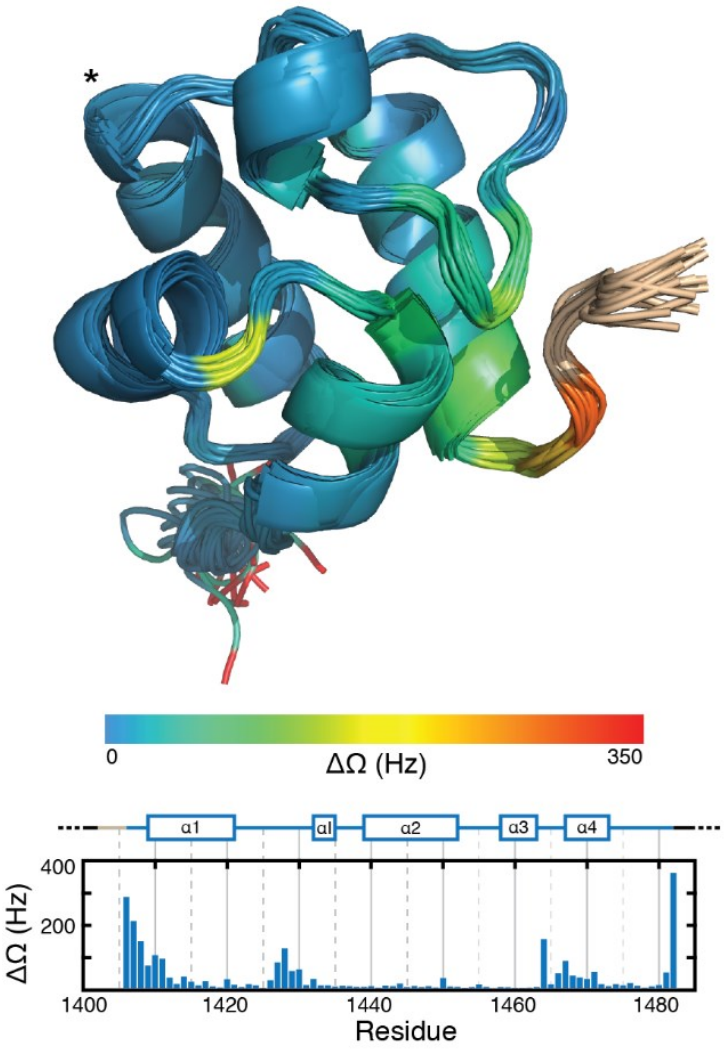

Figure 5.4 CSPs between the full length and truncated constructs. The perturbations are largely limited to contact points between loop0 and the core, which therefore maintains the same fold. Orientation as in Figure 5.1 (A, C)

5.2 Influence of loop0 on the dynamics of $PCP1_{ybt}$

5.2.1 Slow time-scale dynamics of $PCP1_{ybt}$

We used ^{15}N Chemical Exchange Saturation Transfer (CEST) experiments[60] to probe for secondary conformational states in $PCP1_{ybt}$. ^{15}N CEST experiments have emerged as a method to characterize slow time-scale exchange

processes in proteins. By scanning the ^{15}N spectral width with a saturating B_1 field, CEST experiments can reveal the existence of "invisible" signals arising from secondary conformational states through their influence on the signals of the primary conformation.

CEST profiles from residues throughout both constructs of $PCP1_{ybt}$ displayed clear signs of two-state (A & B) exchange. The CEST profiles were fit to a model of a two-state exchange process, and various parameters were extracted, including the chemical shift of each conformational state (ω_A & ω_B) as well as the exchange parameters k_{ex} and p_B. The value p_B represents the population of state B, the minor state, whereas k_{ex} represents the sum of the rate constants of the exchange process, i.e. $k_{ex} = k_{AB} + k_{BA}$.

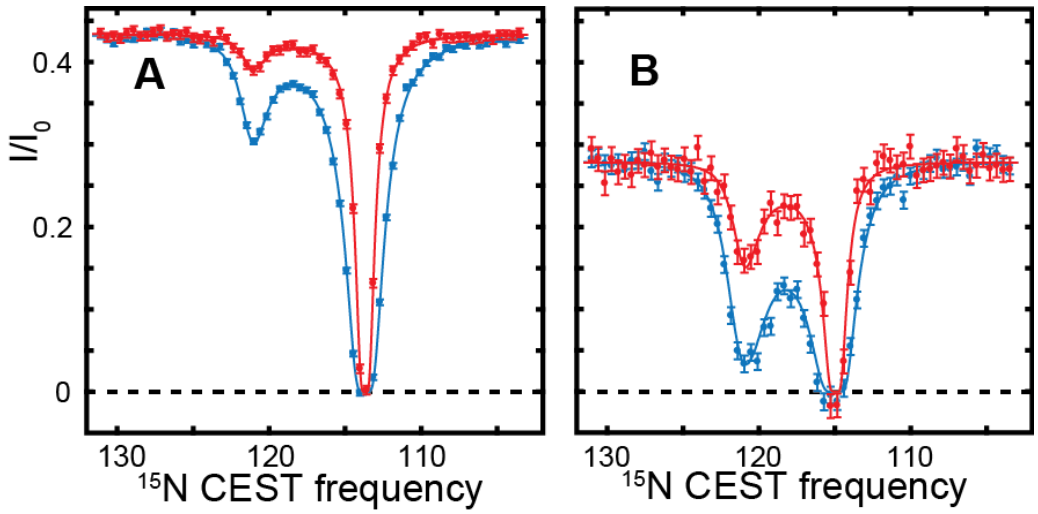

Figure 5.5 CEST reveals unfolded states. (A, B) CEST profiles of residue E1429 using nominal B_1 field strengths of 12.5 Hz (red) and 25 Hz (blue) for the full-length (A) and short (B) constructs.

The fitting strategy employed was akin to that performed previously[60]. First, CEST profiles for all residues of the core were individually fit. Residues in the N- and C-terminal disordered regions did not display signals of secondary

conformers. The core residues for full length $PCP1_{ybt}$ were defined as 1403-1478, and the core residues for the truncated construct were defined as 1406-1478. In this fit, the chemical shift for the major state conformation (ω_A) was held fixed to its initial value while all other fitting parameters were allowed to vary.

In cases where the difference in chemical shift between the major and minor state signals ($\Delta\omega = \omega_B - \omega_A$) is too small, the exchange parameters cannot be accurately fit. In the limit where $\Delta\omega$ approaches zero, all possible values of the exchange parameters result in identical spectra. Thus, to accurately extract the exchange parameters, we must exclude residues with a value of $\Delta\omega$ below a particular threshold. This threshold was established individually for each construct given examination of the CEST profiles and observation of distinct clustering of the exchange parameters above the threshold (Figure 5.6 (a, d)). In the full length construct, CEST profiles with $\Delta\omega > 2.0$ PPM could be identified as possessing a secondary signal, and their individual fits showed distinct clustering in the scatter plot of k_{ex} and p_B. In the truncated construct, the threshold was reduced to $\Delta\omega >$ 1.5 PPM. The threshold could be reduced for this construct because of the increased amplitude of the minor state signals with respect to those of the major state. This fact is reflected in the larger population of the minor state in this construct (see below).

Next, a second fit was performed to extract the common exchange parameters. Only residues with $\Delta\omega$ above the threshold were included in this fit. A total of 38 residues were included in this fit for the full length construct while 40 residues were included for the truncated construct. Here, the exchange

parameters k_{ex} and p_B were fit globally for all of the residues, and the chemical shift of the major state signal (ω_A) was relieved of its constraint and allowed to vary.

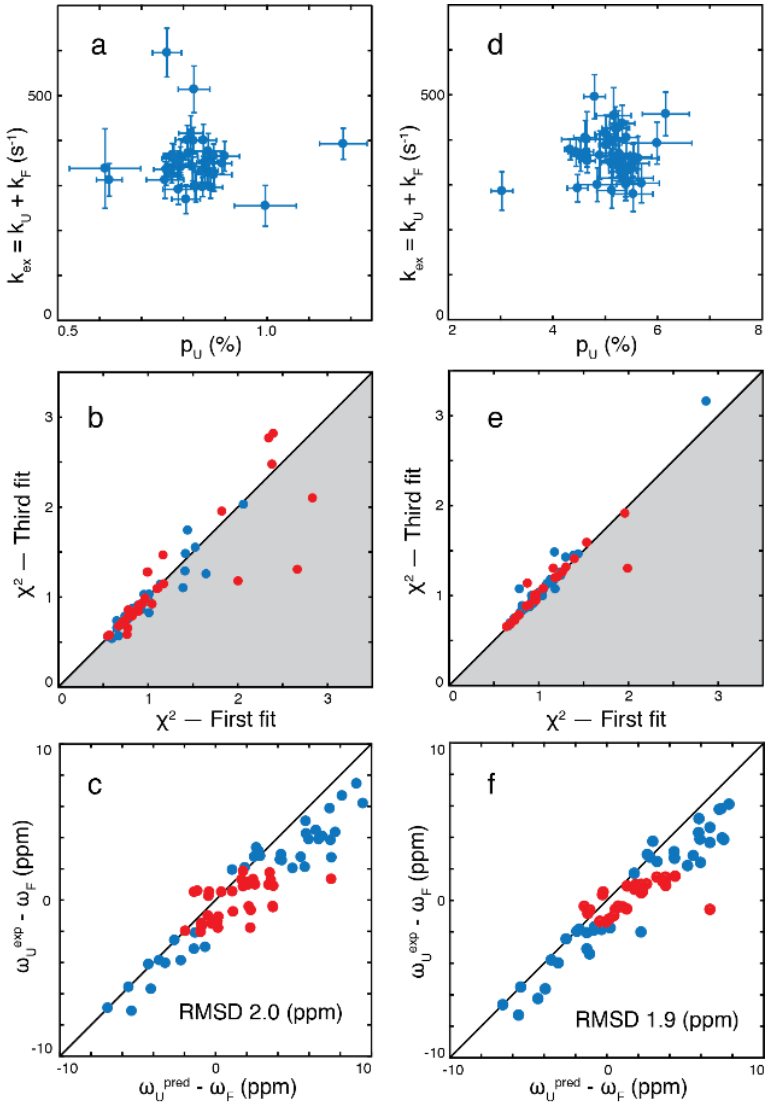

Figure 5.6 ^{15}N CEST results for the full length (a-c) and truncated (d-f) constructs. (a, d) Scatter plot of the exchange parameters k_{ex} and p_U (= p_B) after the first fit. (b, e) Scatter plot of each residue's χ^2 value from the first and third fits where k_{ex} and p_B (p_U) are held constant. Points below the diagonal line, in the gray, shaded region, improved their χ^2 value. Residues shown in red were excluded from the second fit. (c, f) Scatter plot of the experimental values of $\Delta\omega$ and those predicted based on a transition to a random coil. Residues shown in red were excluded from the second fit but included in the third fit. (a) Of the six outlier residues observed in this plot, the χ^2 values in five of them are improved from the first to the third fit. The χ^2 of the sixth residue is increased only slightly, from 1.44 to 1.74. (b) Four residues are not shown in this plot because of their large values of χ^2. All four improve from the first to the third fit. (d) The χ^2 value of the single outlier in this plot increased slightly, from 1.18 in the first fit to 1.48 in the third fit.

Finally, we used a third fit of all the core residues to confirm that the CEST profiles observed for residues with Δω below the threshold were still consistent with the exchange parameters derived from the second fit. In this fit, we held the exchange parameters k_{ex} and p_B fixed to the values obtained in the second fit while again allowing the chemical shift of the major state confirmation (ω_A) to vary. Indeed, the value of χ^2 for every residue either decreased from the first to the third fit or rose only slightly (Figure 5.6 (b, e)), indicating that the observed secondary signals in all residues originate from a common exchange process.

In both constructs, the minor conformation was determined to be an unfolded-like state because of correlation between the observed change in chemical shift and that predicted for a transition to a random-coil (Figure 5.6 (c, f)). Sequence-specific random coil chemical shifts for each residue were obtained from the Neighbor Corrected Intrinsically Disordered Protein Library[101] based on formulas described previously[101]. Furthermore, excellent agreement between the chemical shifts of the minor state in each construct indicate that both constructs unfold to the same random-coil state (Figure 5.7). Thus, moving forward we relabeled the conformational states from A and B to F and U for folded and unfolded respectively.

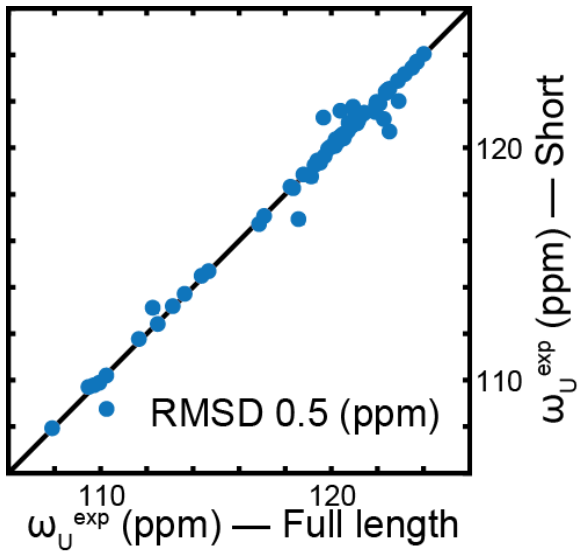

Figure 5.7 Chemical shifts of the unfolded state ω_U in the full length and short constructs. The same unfolded state is observed in both constructs.

We can conclude from our investigation that truncation of loop0 significantly destabilizes the folded core of $PCP1_{ybt}$. Removal of loop0 increases the population of the unfolded state from 0.825 ± 0.005 % in the full length protein to 4.96 ± 0.04 % in the truncated construct. Furthermore, the similarity of the exchange parameters k_{ex} in the full length (324 ± 6 s^{-1}) and truncated constructs (353 ± 6 s^{-1}) indicates that the difference between the exchange process in the two constructs is primarily due to changes in the rate of unfolding. The rate of folding remains relatively steady at values of approximately 321 s^{-1} and 335 s^{-1} respectively, whereas the rate of unfolding increases from approximately 2.7 s^{-1} to 17.5 s^{-1} upon removal of loop0. Thus, loop0 contributes to the stability of the folded core.

5.2.2 Dynamics of $PCP1_{ybt}$ at µs time-scales

Next, we used CPMG relaxation dispersion (RD) experiments[102] to identify evidence of conformational exchange on faster time-scales. RD experiments use

a series of NMR pulses to refocus magnetization that has been defocused by the effects of conformational exchange. The experiments change the effective R_2 relaxation rate $R_{2,eff}$ as a function of the pulsing frequency. At fast enough pulsing frequencies, $R_{2,eff}$ will decay to the baseline relaxation rate R_2. The difference between $R_{2,eff}$ and R_2 is known as R_{ex}, the contribution to relaxation from exchange. Fitting the R_{ex} dispersion profile allows extraction of some of the exchange parameters. When compared to CEST experiments, RD experiments provide less direct evidence of exchange at millisecond time-scales but can characterize faster exchange processes, up to microsecond time-scales.

Investigation of structural fluctuations in the folded $PCP1_{ybt}$ core using RD did not reveal conformational exchange processes beyond the aforementioned unfolding. Residues throughout the core of the full length protein showed evidence of dispersion (Figure 5.8 (a, b)), with maximum values of R_{ex} in the range 2-3 s^{-1} compared to baseline values of R_2 in the range 12-13 s^{-1}. Given knowledge of the unfolding exchange process identified by ^{15}N CEST, we sought to determine the contribution of unfolding to the observed relaxation dispersion. To do so, we fit all residues displaying dispersion with a mathematical model assuming slow exchange[103]. Comparison of the exchange parameters derived from relaxation dispersion to those obtained by ^{15}N CEST reveals good agreement (Figure 5.8 (c, d)). Thus, the unfolding process largely accounts for the observed dispersion and other exchange processes cannot be identified within the framework of our experimental conditions. If other exchange processes exist, they are masked by unfolding.

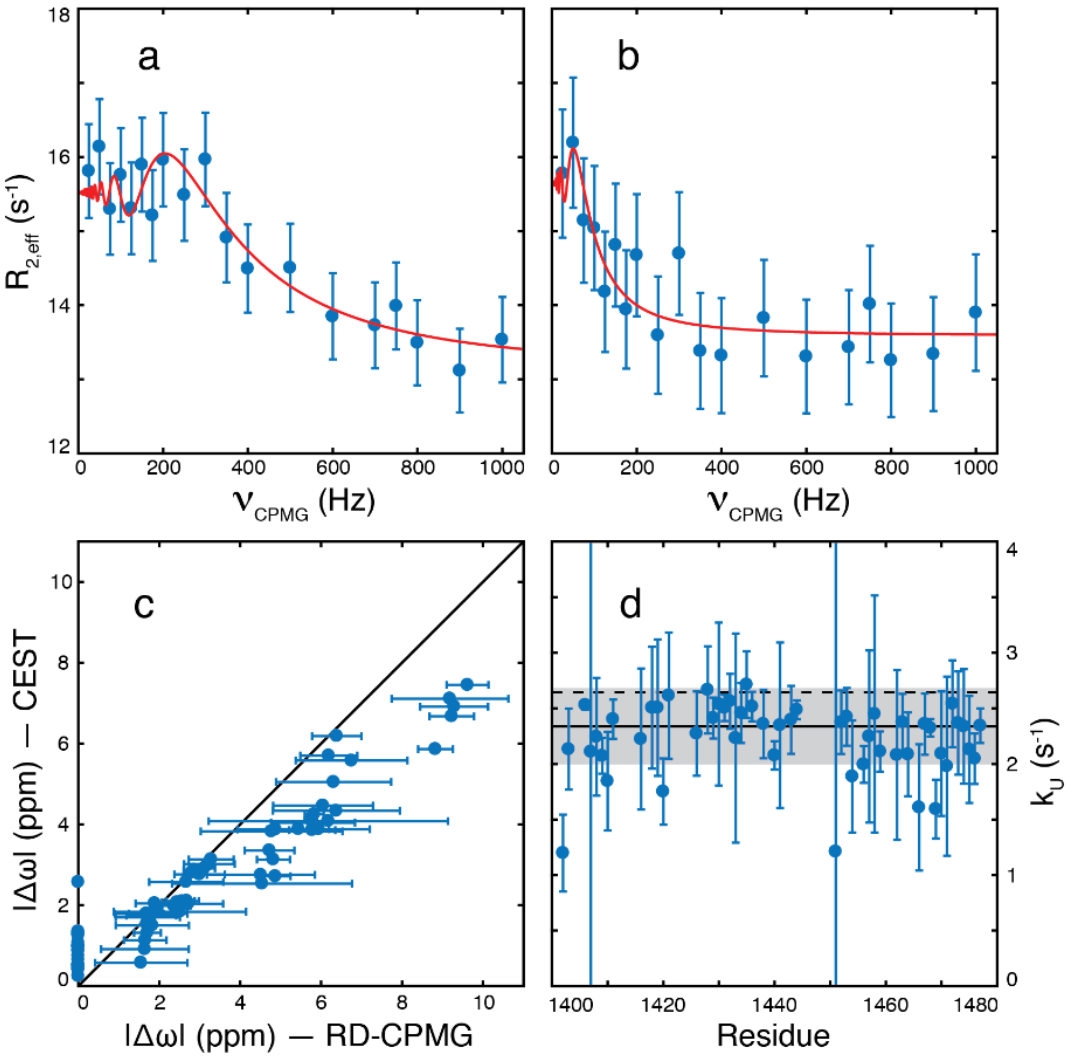

Figure 5.8 Imposing the slow exchange relaxation dispersion model reveals agreement with the results of ^{15}N CEST. (a) The relaxation dispersion profile of residue E1429 and its associated fit to the slow exchange model (b) The relaxation dispersion profile of residue L1440 and its associated fit to the slow exchange model (c) Scatter plot of the absolute value of the chemical shift difference Δω between the two states as measured by RD and CEST. The slow exchange RD model is not able to discern the sign of Δω. (d) Residue-specific plot of the unfolding rate k_U as measured by RD. The median value of k_U is indicated by the solid horizontal line. The shaded area represents the RMSD of the values of k_U with respect to the median. The dashed line represents the value of k_U as measured by ^{15}N CEST.

5.2.3 Probing PCP1$_{ybt}$ dynamics at ps-ns time-scales

To probe for allosteric communication between loop0 and the protein core at faster time-scales (ps-ns), we also performed Model Free analysis of relaxation

data from the truncated form of $PCP1_{ybt}$. The Model Free formalism provides a means to quantify fast time-scale motions in proteins, typically through their effects on the NMR relaxation properties of backbone ^{15}N nuclei. Because motions on the ps-ns time scale are faster than the tumbling of the molecule in solution, successful quantification of such motions is critically dependent on accurate characterization of the molecule's rotational diffusion. A rotational diffusion tensor describes the molecule's tumbling in solution, and the corresponding rotational correlation time (τ_c) is a simple quantity that characterizes the rate of diffusion. Longer values of τ_c indicate slower molecular tumbling.

The experimentally determined rotational correlation time for the truncated construct, 4.8 ns, falls exactly in line with the value predicted from its molecular weight, also 4.8 ns. Such values are predicted based on an assumption of Brownian rotational diffusion and a spherical approximation for the protein. These assumptions seem reasonable for the compact, folded core of the truncated construct. However, we should not expect the spherical assumption to hold true for the full length construct which includes extensive, uncompact linker residues.

Indeed, we note that the full length construct tumbles 30% more slowly than anticipated. Its predicted value of τ_c was 6.6 ns while it's experimentally determined value of τ_c was 8.5 ns. This behavior is presumably caused by the flexible regions of the linkers, in agreement with a study on such effects[104]. Furthermore, size exclusion chromatography (SEC) indicates that full length $PCP1_{ybt}$ has an unusually large hydrodynamic volume (Figure 5.9), yet multi-angle light scattering (MALS) reveals a monomeric protein in solution (Figure 5.10).

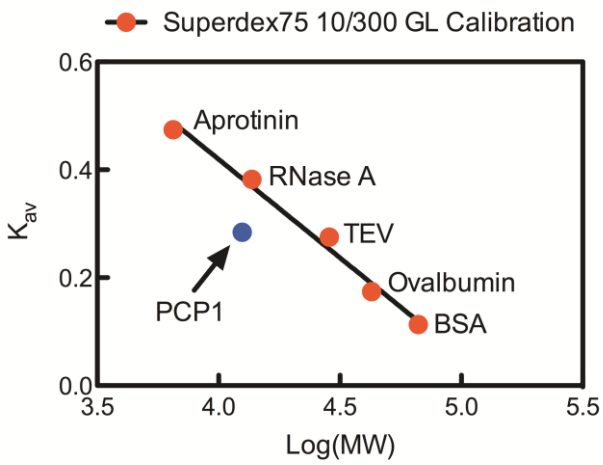

Figure 5.9 Full length $PCP1_{ybt}$ elutes at a large molecular weight during SEC. Based on its gel phase distribution coefficient (K_{av}), its predicted molecular weight (MW) is 22.9 kDa.

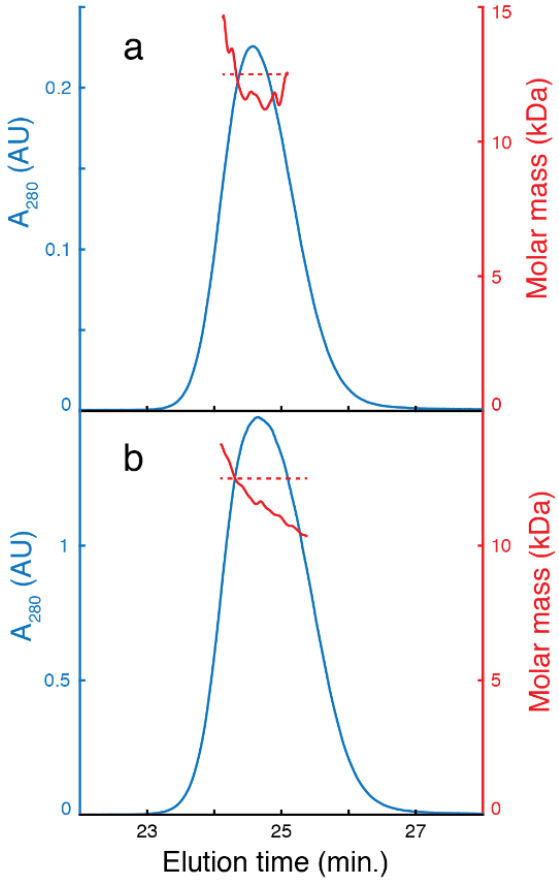

Figure 5.10 SEC-MALS traces for two samples of full length $PCP1_{ybt}$ (760 µM, ^{15}N labeled). The dashed line in both plots indicates the anticipated molar mass of ^{15}N labeled protein, 12.496 kDa. (a) 20 µL injection (b) 200 µL injection

Thus, to prevent confounding hydrodynamic effects when interpreting changes between constructs, we prepared a third construct of $PCP1_{ybt}$. This construct is identical to the short construct discussed throughout the text except that loop0 has been restored by extending the linker by four residues (beige in the secondary structure diagrams throughout the chapter). Model Free analysis of the new construct confirms that its rotational correlation time of 5.2 ns agrees well with the prediction of 5.1 ns. As a result, comparisons of Model Free parameters between the two short constructs will faithfully report on modulations in fast time-scale dynamics due only to contact between loop0 and the protein core. Plots of the ^{15}N relaxation data and modeling parameters S^2 and R_{ex} for each of the three constructs are shown in Figure 5.11, Figure 5.12 and Figure 5.13.

As mentioned previously, the Model Free parameter S^2 broadly characterizes the extent of motions at a particular site that are faster than the rotational diffusion of the molecule. Values of S^2 near one indicate relative rigidity with respect to the reference frame defined by the rotational diffusion tensor, whereas values near zero indicate extensive motions. It is particularly important to note, however, that changes in S^2 cannot be strictly interpreted as an increase or decrease in flexibility. The value of S^2 depends not only on the extent of flexibility but also on the type of motion and its angles relative to the reference frame.

On the other hand, the value of R_{ex} characterizes motions on time-scales slower than rotational diffusion of the molecule. Proper modeling of ps-ns motions using the Model Free formalism requires measurement of the baseline value of R_2, uncorrupted by the influence of any slower conformational exchange events.

Consequently, R_2 measurements for this purpose are typically performed with strong CPMG pulsing to diminish residual contributions from R_{ex}. However, complete removal is not always possible, and Model Free fitting software usually compensates by directly fitting any lingering R_{ex}. This is particularly relevant in the presence of conformational exchange events occurring on time-scales too fast to be probed by relaxation dispersion yet still slower than molecular tumbling. In such cases, there may be substantial residual R_{ex}.

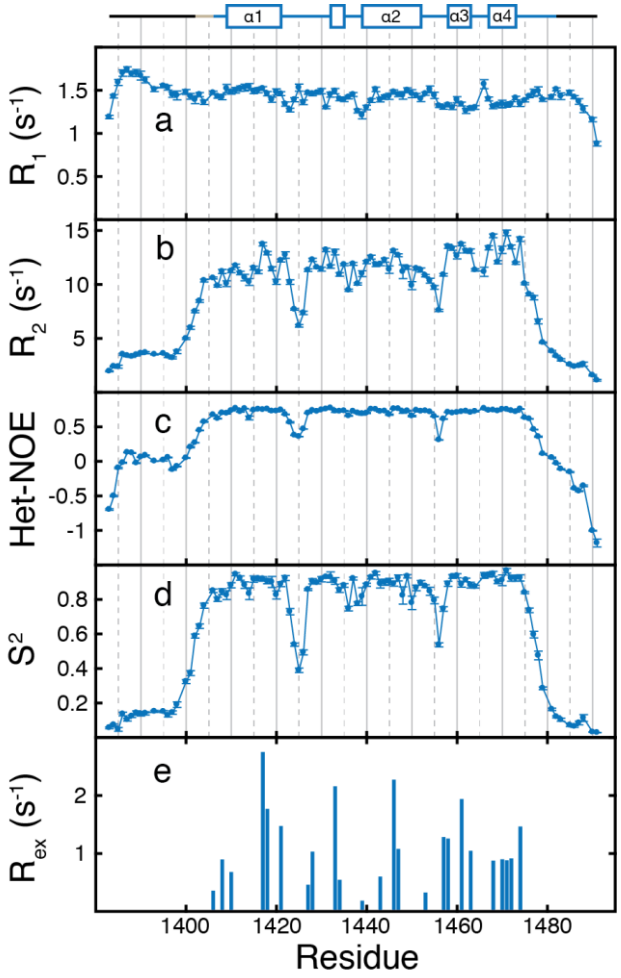

Figure 5.11 Relaxation parameters for full length $PCP1_{ybt}$. The secondary structure is illustrated above the plots. (a) R_1 relaxation rates. (b) R_2 relaxation rates. (c) Heteronuclear NOE (d) Order parameter S^2 (e) R_{ex}, the residual contribution to R_2 from conformational exchange

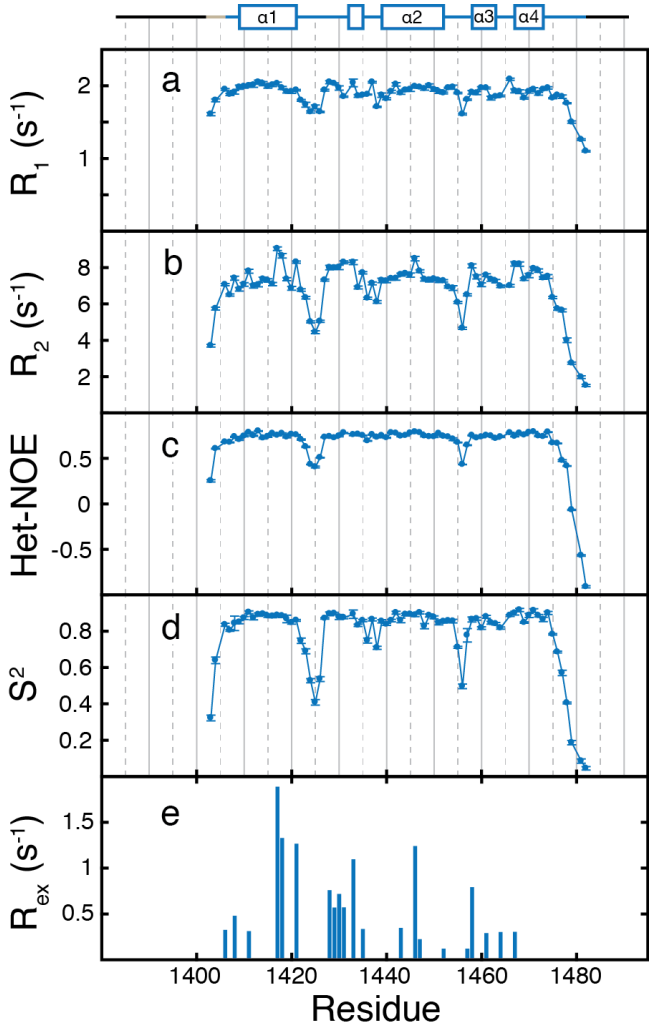

Figure 5.12 Relaxation parameters for truncated $PCP1_{ybt}$ with loop0. The secondary structure is illustrated above the plots. (a) R_1 relaxation rates. (b) R_2 relaxation rates. (c) Heteronuclear NOE (d) Order parameter S^2 (e) R_{ex}, the residual contribution to R_2 from conformational exchange

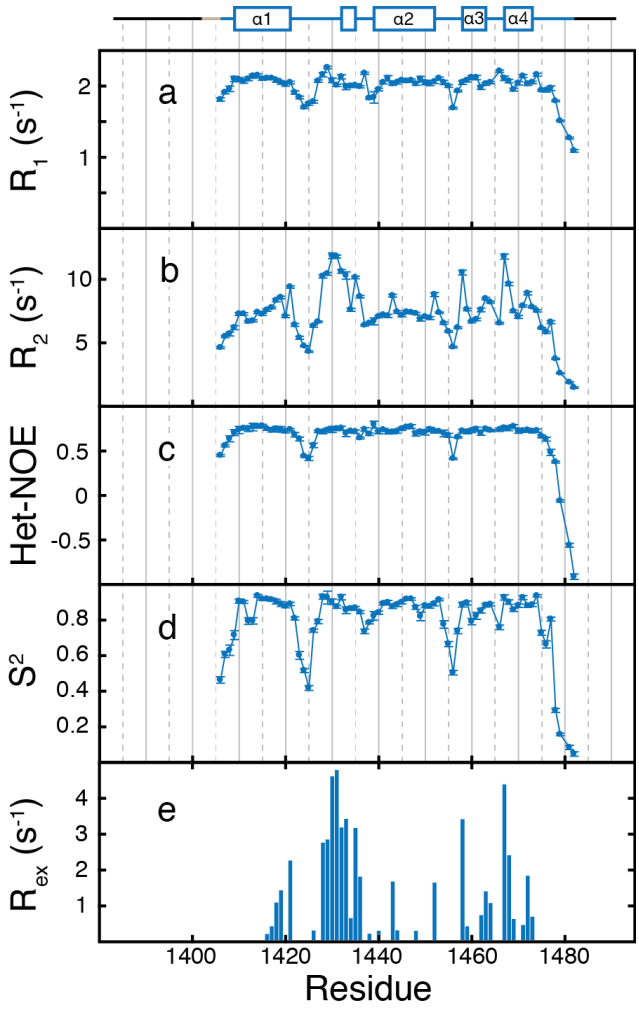

Figure 5.13 Relaxation parameters for fully truncated $PCP1_{ybt}$. The secondary structure is illustrated above the plots. (a) R_1 relaxation rates. (b) R_2 relaxation rates. (c) Heteronuclear NOE (d) Order parameter S^2 (e) R_{ex}, the residual contribution to R_2 from conformational exchange

Several differences emerge when evaluating the change in S^2 (ΔS^2) between the two short constructs. Figure 5.14 maps the absolute value of ΔS^2 ($|\Delta S^2|$) to the ribbon width of the $PCP1_{ybt}$ structure. Using $|\Delta S^2|$ emphasizes that changes in S^2 do not necessarily represent an increase or decrease in flexibility.

When loop 0 is shortened, we observe large changes in S^2 at the site of truncation and at the contact points between loop 0 and the core. However,

additional sites throughout the protein appear to be affected as well. Changes in S^2 within loop3 may reflect the subtle rearrangements invoked to explain the CSP at H1464. Changes in S^2 in the C-terminal tail contacting loop2 denote communication between both ends of the protein, which may occur indirectly through α3 and α4 via loop3, which contacts loop0. Further changes in S^2 are apparent at the N-terminus of loop1, which has been implicated in binding between related aryl carrier proteins and their associated adenylation domains[105]. Finally, the region immediately preceding the post-translationally modified serine, residues 1436-1438, is of particular interest. The S^2 of G1437 is substantially larger than that of its neighbors when loop0 contacts the core (Figure 5.12 (d)) but the pattern is inverted in the shorter construct (Figure 5.13 (d)), denoting a substantial change in the amplitude and/or orientation of the motion. In any case, the conformational landscape in this region is modified, likely impacting communication between $PCP1_{ybt}$ and other, catalytic domains during synthesis. Critically, this dynamic region belongs to the major conserved motif of carrier proteins[106].

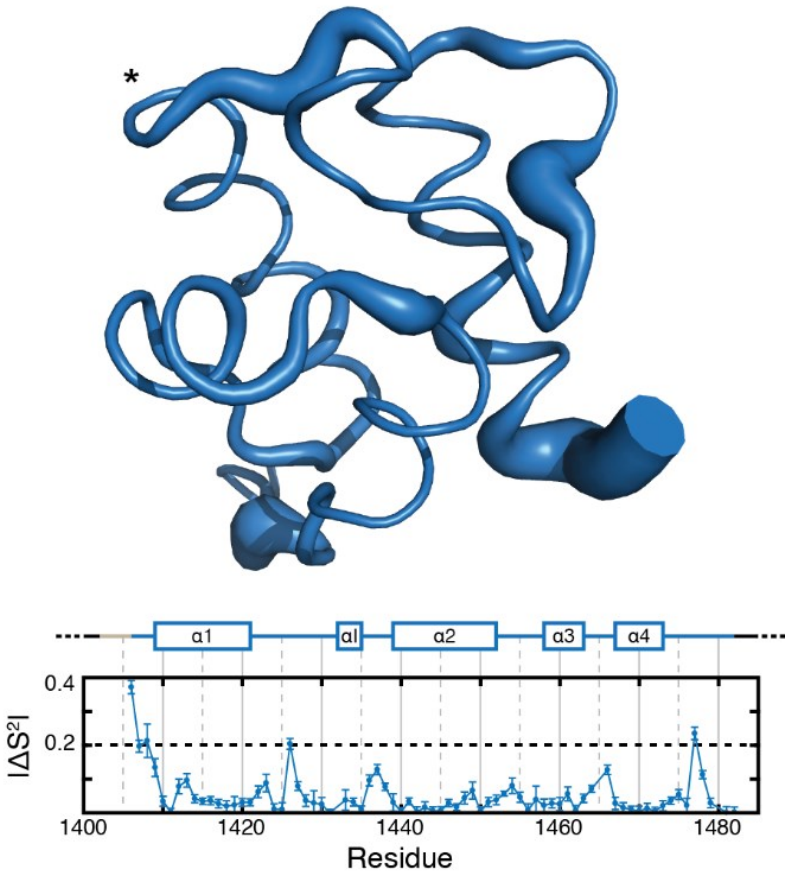

Figure 5.14 Modulation of fast (ps-ns) dynamics in $PCP1_{ybt}$ by loop0. A thicker width denotess a larger absolute change in S^2 ($|\Delta S^2|$) between constructs when truncating loop0, up to a cap defined by the dashed line. The residues of loop0 that are removed are represented by the beige region in the secondary structure diagram.

In addition to the differences in S^2 between constructs, we also note the elevated values of R_{ex} present throughout the fully truncated form (Figure 5.13). Rather than being indicative of a new conformational exchange process too fast to be measured by relaxation dispersion, we instead suspected that this result primarily reflected residual contributions from the strong unfolding exchange process present in the fully truncated construct. To validate this claim, we back calculated the anticipated residual R_{ex} at the pulsing frequency (ν_{CPMG}) used in our R_2 experiment based on the exchange parameters measured by ^{15}N CEST. We

used the CR72 model[107] to calculate R_{ex} at the value of ν_{CPMG} = 758 Hz. A comparison of calculated and fit R_{ex} values is presented in Figure 5.15. In general, the residues with large values of R_{ex} fit during Model Free analysis are also those expected to have a large R_{ex} based on the ^{15}N CEST data. As a result, we conclude that any additional exchange processes, if they exist, are again masked by contributions to R_{ex} from the slow-exchange unfolding.

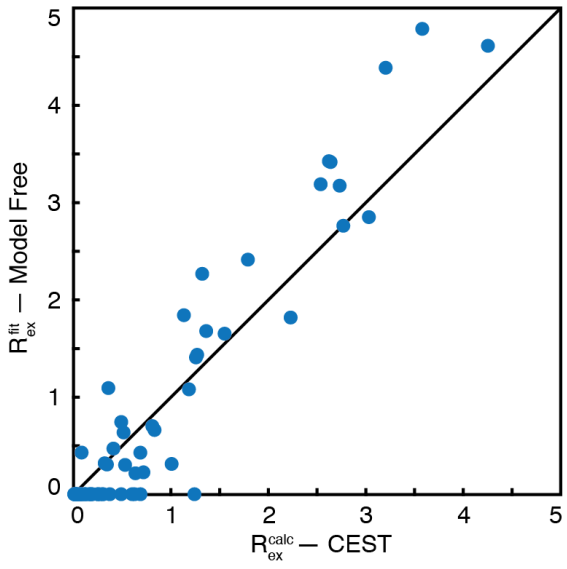

Figure 5.15 Residual contribution of R_{ex} in R_2 relaxation experiments performed at ν_{CPMG} = 758 Hz for fully truncated $PCP1_{ybt}$. Experimental values were fit during Model Free analysis. Predicted values were determined based on the exchange parameters measured by ^{15}N CEST. The R_{ex} values were calculated using equations described previously[107].

6 Discussion

6.1 NMR assignment with 4D covariance correlation maps

In chapter 3, we introduced new pre- and post-processing methods to enhance the utility of covariance NMR spectra. We showed that applying a spectral derivative along the subsumed dimension prior to performing unsymmetric covariance calculations reduces or eliminates the presence of unwanted false positive artifacts. Additionally, we demonstrated that several covariance NMR spectra connecting the same set of nuclei through different common nuclei could be combined with element-wise multiplication to further reduce the frequency of false positive artifacts. Finally, we presented novel four-dimensional covariance correlation maps that harness the information of conventional 3D assignment spectra and combine them into a single 4D spectrum, and we developed an easy to use processing script to facilitate their calculation.

Overall, we believe 4D covariance spectra produced with our methods to be an important addition to the procedures employed during NMR resonance assignment that should help promulgate studies of systems with increased spectroscopic complexity. Future work in this field might integrate our covariance processing techniques into existing NMR assignment software applications. Such efforts would ease the burden on researchers and promote more widespread application of the technique. In addition, covariance NMR techniques could also be incorporated into automated assignment procedures to further enhance their accuracy in and applicability to proteins with crowded spectra.

6.2 Measuring NMR relaxation rates with accordion spectroscopy

In chapter 4, we outlined the development of a graphical, user-friendly software package SARA for processing data from accordion relaxation experiments. It harnesses the development speed and accessibility of the MATLAB environment to bring accordion data analysis to a wide audience. In order to provide users with the tools necessary to investigate a wide array of proteins, we implemented two analysis methods in SARA that span the majority of approaches presented in the literature so far. One method offers the highest possible fitting precision, but it may be applied sub-optimally without close inspection by the user. As an alternative, we have developed a new method which harnesses the ability of the Fourier transform to separate signals prior to analysis. We provided guidelines for its appropriate application and discussed the limitations of its use. We also included a symmetrization feature that may be used both to identify cases of strongly overlapped NMR signals and resolve cases of slight overlap. The strength of SARA lies in its ability to evaluate and fit data from multiple perspectives. To harness this strength, we suggested a protocol relying on both analysis procedures that together ensures a reliable and precise estimation of relaxation rates using a robust and interactive software environment.

The accordion method is an efficient way to measure relaxation rates in a fraction of the time needed using traditional experiments. We hope that the tools we have designed for analyzing relaxation rates will facilitate and promote routine application of the accordion method during biological studies and thereby

encourage a more systematic investigation of protein dynamics by NMR. Future work in this field would evaluate the circumstances under which the accordion method is superior to traditional NMR relaxation experiments and would promote the use of accordion spectroscopy in these cases.

6.3 Molecular cross-talk between an NRPS carrier protein and its linkers

In chapter 5, we determined the solution structure of $PCP1_{ybt}$, an NRPS peptidyl carrier protein from Yersiniabactin Synthetase, and we identified contacts between its linker residues and its folded core. We found that a region of the N-terminal linker, loop0, provides a six-fold increase in fold stability and influences the ps-ns time-scale dynamics of residues throughout the protein, particularly those directly preceding the post-translationally modified serine, which are conserved in sequence and structure. This region has been shown to be dynamic in other carrier proteins, and it is affected by post-translational modifications[94]. Loop0 may act as an allosteric sensor between the N-terminal linker and the protein core, propagating molecular events affecting the core to linkers in a manner that ensures their remodeling during sequential NRPS domain interactions. These findings underline the importance of unstructured regions in multi-domain proteins, and highlight the role of protein dynamics in molecular communication.

Future investigations of the role played by protein dynamics in $PCP1_{ybt}$ would interrogate the changes in dynamics of the protein upon post-translational modification to its holo and loaded states. Furthermore, biochemical and structural

characterization of loop0 mutants in the full synthetase would confirm the influence of the interaction between loop0 and the $PCP1_{ybt}$ core on the dynamic interactions that occur between domains during synthesis.

6.4 Conclusions

In this work, I have presented contributions to the field of NMR spectroscopy that alleviate some of the difficulties faced by investigators when working with challenging proteins. Crowded spectra from large, disordered and some α-helical proteins are often difficult to assign. In such scenarios, rather than rely on peak picking – an ambiguous and unreliable abstraction of the underlying data – the covariance NMR techniques introduced herein will help to facilitate assignment by mathematically combining NMR spectra to derive assignment candidates instead. The covariance pre- and post-processing steps I have introduced are instrumental in reducing the prevalence of artifacts in covariance spectra. These artifacts have presented a major barrier to broad application of covariance NMR methods, and consequently, artifact suppression will help to increase the utility of the technique. Furthermore, the processing steps are general and can be implemented in a wide range of strategies employing covariance NMR. They are not limited to covariance between 3D spectra, rather they can be applied to arbitrary combinations of spectra to create any conceivably useful set of NMR correlations. Future applications of covariance NMR methods may include automated assignment algorithms, enhanced hyperdimensional NMR spectroscopy and other high-order, NMR spectra techniques, as well as possible uses in solid state and small molecule NMR analysis.

The software package SARA, developed for the analysis of data acquired with accordion relaxation spectroscopy, will improve the ability of researchers to collect and analyze NMR dynamics data on expensive and unstable samples. Accordion NMR methods reduce the minimum time required to measure NMR relaxation rates. This is especially helpful in cases where a wide range of rates is expected, for example when a protein contains disordered and/or strongly exchanging residues. Extension of accordion methods to 3D NMR relaxation experiments may permit NMR dynamics studies of larger proteins, because spectral crowding is reduced without a concomitant increase in acquisition time. Furthermore, future applications of accordion NMR spectroscopy to the measurement of cross-correlated relaxation rates may provide new avenues to study protein dynamics, especially when sample stability is a major concern. My work on SARA provides a framework to carefully analyze the results of accordion experiments in a manner that ensures the highest quality relaxation rates are derived from the data. Consequently, the approaches I have described will be instrumental in expanding the prevalence and utility of accordion NMR techniques.

Finally, I have applied NMR experiments to study the dynamics of a protein domain from an important class of enzymatic systems responsible for the synthesis of a number of pharmaceutically relevant molecules. As discussed previously, non-ribosomal peptide synthetases use a modular, assembly-line architecture to synthesize secondary metabolites that have found many medicinal uses. The prospect of successfully re-engineering NRPS systems to create arbitrary peptides remains an enticing goal. However, approaches based on domain swapping

between synthetases have had little success thus far. My work revealed that a portion of the N-terminal linker region in an NRPS peptidyl carrier protein makes contact with the folded core of the domain and allosterically modulates its dynamics. Future studies with the protein may elucidate the dynamic mechanisms by which contact between linker and the core affects the site of post-translational modification, for example by mutational analysis of the residues putatively responsible for propagation. Furthermore, my results will help define the proper sites of excision when extracting carrier proteins from these megadalton enzymatic assemblies and may shed light on the modes of rotation and translation of carrier proteins during synthesis. In particular, it is known that carrier proteins are often guided by rotations occurring in adjacent adenylation domains. Further study of this linker region by NMR in adynelation/carrier protein di-domain constructs may provide a stronger understanding of the role played by the linker residues connecting these two domains and would improve our ability to modify and direct these interactions.

In total, my work represents a contribution to the field of structural biology intended to help better assess the dynamic nature of proteins with NMR spectroscopy. I have developed data analysis techniques and software to broaden the applicability of two under-utilized NMR techniques, and I have demonstrated the utility of NMR experiments to study the complex and dynamic interactions that occur within proteins. My results have laid the foundation for future work that will deepen our understanding of proteins and enhance our ability to engineer them.

7 References

1. Eisenmesser EZ, Millet O, Labeikovsky W, et al. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*. 2005;438(7064):117-121. doi:10.1038/nature04105.

2. Henzler-wildman KA, Thai V, Lei M, et al. Intrinsic motions along an enzymatic reaction trajectory. 2007;450(December). doi:10.1038/nature06410.

3. Wand AJ. The dark energy of proteins comes to light : conformational entropy and its role in protein function revealed by NMR relaxation. *Curr Opin Struct Biol*. 2013;23(1):75-81. doi:10.1016/j.sbi.2012.11.005.

4. Akke M, Bruschweiler R, Palmer AG, Iii J. NMR Order Parameters and Free Energy: An Analytical Approach and Its Application to Cooperative Ca2+ Binding by Calbindin D9k. *J Am Chem Soc*. 1993;115(21):9832-9833.

5. Sugase K, Dyson HJ, Wright PE. Mechanism of coupled folding and binding of an intrinsically disordered protein. 2007;447(June). doi:10.1038/nature05858.

6. Korzhnev DM, Kay LE. Probing invisible, low-populated States of protein molecules by relaxation dispersion NMR spectroscopy: an application to protein folding. *Acc Chem Res*. 2008;41(3):442-451. doi:10.1021/ar700189y.

7. Harden BJ, Mishra SH, Frueh DP. Effortless assignment with 4D covariance sequential correlation maps. *J Magn Reson*. 2015;260:83-88. doi:10.1016/j.jmr.2015.09.007.

8. Harden BJ, Frueh DP. Covariance NMR processing and analysis for protein assignment. *Methods Mol Biol [submitted]*. 2016.

9. Sprangers R, Kay LE. Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature*. 2007;445(7128):618-622. doi:10.1038/nature05512.

10. Tugarinov V, Kay LE. Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *J Am Chem Soc*. 2003;125(45):13868-13878. doi:10.1021/ja030345s.

11. Harden B, Frueh D. SARA: a software environment for the analysis of relaxation data acquired with accordion spectroscopy. *J Biomol NMR*. 2014;58(2):83-99. doi:10.1007/s10858-013-9807-x.

12. Bodenhausen G, Ernst R. The accordion experiment, a simple approach to three-dimensional NMR spectroscopy. *J Magn Reson*. 1981;45(2):367-373. doi:10.1016/0022-2364(81)90137-2.

13. Mandel a. M, Palmer a. G. Measurement of Relaxation-Rate Constants Using Constant-Time Accordion NMR Spectroscopy. *J Magn Reson Ser A*. 1994;110(1):62-72. doi:10.1006/jmra.1994.1182.

14. Rabier P, Kieffer B, Koehl P, Lefèvre J-F. Fast measurement of heteronuclear relaxation: frequency-domain analysis of NMR accordion spectroscopy. *Magn Reson Chem*. 2001;39(8):447-456. doi:10.1002/mrc.870.

15. Chen K, Tjandra N. Direct measurements of protein backbone 15N spin relaxation rates from peak line-width using a fully-relaxed Accordion 3D HNCO experiment. *J Magn Reson*. 2009;197(1):71-76. doi:10.1016/j.jmr.2008.12.001.

16. Carr P a, Fearing D a, Palmer a G. 3D accordion spectroscopy for measuring 15N and 13CO relaxation rates in poorly resolved NMR spectra. *J Magn Reson*. 1998;132(1):25-33. doi:10.1006/jmre.1998.1374.

17. Guenneugues M, Gilquin B, Wolff N, Ménez A, Zinn-Justin S. Internal motion time scales of a small, highly stable and disulfide-rich protein: a 15N, 13C NMR and molecular dynamics study. *J Biomol NMR*. 1999;14(1):47-66.

18. The MathWorks Inc. MATLAB. 2014.

19. Klein E, Smith DL, Laxminarayan R. Hospitalizations and Deaths Caused by Methicillin-Resistant Staphylococcus aureus , United States, 1999-2005. *Emerg Infect Dis*. 2007;13(12):1999-2005.

20. Howden BP, Davies JK, Johnson PDR, Stinear TP, Grayson ML. Reduced vancomycin susceptibility in Staphylococcus aureus, including vancomycin-intermediate and heterogeneous vancomycin-intermediate strains: resistance mechanisms, laboratory detection, and clinical implications. *Clin Microbiol Rev*. 2010;23(1):99-139. doi:10.1128/CMR.00042-09.

21. Stevens B, Joska TM, Anderson AC. Progress toward re-engineering non-ribosomal peptide synthetase proteins: a potential new source of pharmacological agents. *Drug Dev …*. 2005;66(August 2005):9-18. doi:10.1002/ddr.

22. Marahiel MA, Stachelhaus T, Mootz HD. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem Rev*. 1997;97(7):2651-2674. doi:10.1016/S0006-3495(02)75175-8.

23. Tanovic A, Samel S a, Essen L-O, Marahiel M a. Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science*. 2008;321(5889):659-663. doi:10.1126/science.1159850.

24. Mitchell C a, Shi C, Aldrich CC, Gulick AM. Structure of PA1221, a nonribosomal peptide synthetase containing adenylation and peptidyl carrier protein domains. *Biochemistry*. 2012;51(15):3252-3263. doi:10.1021/bi300112e.

25. Marahiel M a, Essen L-O. *Nonribosomal Peptide Synthetases Mechanistic and Structural Aspects of Essential Domains.* Vol 458. 1st ed. Elsevier Inc.; 2009. doi:10.1016/S0076-6879(09)04813-7.

26. Samel S a, Schoenafinger G, Knappe T a, Marahiel M a, Essen L-O. Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. *Structure*. 2007;15(7):781-792. doi:10.1016/j.str.2007.05.008.

27. Frueh D, Arthanari H, Koglin A, Vosburg D. Dynamic thiolation – thioesterase structure of a non-ribosomal peptide synthetase. *Nature*. 2008;454(August). doi:10.1038/nature07162.

28. Koglin A, Löhr F, Bernhard F, et al. Structural basis for the selectivity of the external thioesterase of the surfactin synthetase. *Nature*. 2008;454(7206):907-911. doi:10.1038/nature07161.

29. Gokhale RS, Khosla C. Role of linkers in communication between protein modules. *Curr Opin Chem Biol*. 2000;4(1):22-27. doi:10.1016/S1367-5931(99)00046-0.

30. Hahn M, Stachelhaus T. Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. *Proc Natl Acad Sci U S A*. 2004;101(44):15585-15590. doi:10.1073/pnas.0404932101.

31. Chiocchini C, Linne U, Stachelhaus T. In Vivo Biocombinatorial Synthesis of Lipopeptides by COM Domain-Mediated Reprogramming of the Surfactin Biosynthetic Complex. *Chem Biol*. 2006;13(8):899-908. doi:10.1016/j.chembiol.2006.06.015.

32. Hahn M, Stachelhaus T. Harnessing the potential of communication-mediating domains for the biocombinatorial synthesis of nonribosomal peptides. *Proc Natl Acad Sci U S A*. 2006;103(2):275-280. doi:10.1073/pnas.0508409103.

33. Gehring a M, DeMoll E, Fetherston JD, et al. Iron acquisition in plague: modular logic in enzymatic biogenesis of yersiniabactin by Yersinia pestis. *Chem Biol*. 1998;5(10):573-586.

34. Carniel E. The Yersinia high-pathogenicity island: an iron-uptake island. *Microbes Infect*. 2001;3(7):561-569.

35. Henderson JP, Crowley JR, Pinkner JS, et al. Quantitative metabolomics reveals an epigenetic blueprint for iron acquisition in uropathogenic Escherichia coli. *PLoS Pathog*. 2009;5(2):e1000305. doi:10.1371/journal.ppat.1000305.

36. Gehring A, Mori I, Perry R, Walsh C. The nonribosomal peptide synthetase HMWP2 forms a thiazoline ring during biogenesis of yersiniabactin, an iron-chelating virulence factor of yersinia pestis. *Biochemistry*.

1998;37(48):17104. doi:10.1021/bi9850524.

37. Keating TA, Miller DA, Walsh CT. Expression, purification, and characterization of HMWP2, a 229 kDa, six domain protein subunit of Yersiniabactin synthetase. *Biochemistry*. 2000;39(16):4729-4739.

38. Keating TA, Suo Z, Ehmann DE, Walsh CT. Selectivity of the yersiniabactin synthetase adenylation domain in the two-step process of amino acid activation and transfer to a holo-carrier protein domain. *Biochemistry*. 2000;39(9):2297-2306.

39. Miller D a, Walsh CT. Yersiniabactin synthetase: probing the recognition of carrier protein domains by the catalytic heterocyclization domains, Cy1 and Cy2, in the chain-initiating HWMP2 subunit. *Biochemistry*. 2001;40(17):5313-5321.

40. Miller DA, Luo L, Hillson N, Keating T a, Walsh CT. Yersiniabactin synthetase: a four-protein assembly line producing the nonribosomal peptide/polyketide hybrid siderophore of Yersinia pestis. *Chem Biol*. 2002;9(3):333-344.

41. Suo Z, Tseng CC, Walsh CT. Purification, priming, and catalytic acylation of carrier protein domains in the polyketide synthase and nonribosomal peptidyl synthetase modules of the HMWP1 subunit of yersiniabactin synthetase. *Proc Natl Acad Sci U S A*. 2001;98(1):99-104. doi:10.1073/pnas.021537498.

42. Suo Z, Walsh C, Miller D. Tandem heterocyclization activity of the multidomain 230 kDa HMWP2 subunit of Yersinia pestis yersiniabactin synthetase: interaction of the 1-1382 and 1383-2035 fragments. *Biochemistry*. 1999;38(42):14023-14035.

43. Suo Z. Thioesterase portability and peptidyl carrier protein swapping in yersiniabactin synthetase from Yersinia pestis. *Biochemistry*. 2005;44(12):4926-4938. doi:10.1021/bi047538s.

44. Mishra SH, Frueh D. Assignment of Methyl NMR Resonances of a 52 kDa Protein with Residue-specific 4D Correlation Maps. *J Biomol NMR*. 2015;62(3):281-290. doi:10.1007/s10858-015-9943-6.

45. Mishra S, Harden B, Frueh D. A 3D time-shared NOESY experiment designed to provide optimal resolution for accurate assignment of NMR distance restraints in large proteins. *J Biomol NMR*. 2014;60(4):265-274. doi:10.1007/s10858-014-9873-8.

46. Hyberts SG, Takeuchi K, Wagner G. Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. *J Am Chem Soc*. 2010;132(7):2145-2147. doi:10.1021/ja908004w.

47. Hyberts SG, Milbradt AG, Wagner AB, Arthanari H, Wagner G. Application

of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson Gap scheduling. *J Biomol NMR*. 2012;52(4):315-327. doi:10.1007/s10858-012-9611-z.

48. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR*. 1995;6(3):277-293.

49. Keller R. *The Computer Aided Resonance Assignment Tutorial*. Goldau, Switzerland: Cantina Verlag; 2004.

50. Sun S, Gill M, Li Y, Huang M, Byrd RA. Efficient and generalized processing of multidimensional NUS NMR data: The NESTA algorithm and comparison of regularization terms. *J Biomol NMR*. 2015;62(1):105-117. doi:10.1007/s10858-015-9923-x.

51. Short T, Alzapiedi L, Brüschweiler R, Snyder D. A covariance NMR toolbox for MATLAB and OCTAVE. *J Magn Reson*. 2011;209(1):75-78. doi:10.1016/j.jmr.2010.11.018.

52. Eaton JW, Bateman D, Hauberg S, Wehbring R. GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations. 2015.

53. Shen Y, Bax A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR*. 2013;56(3):227-241. doi:10.1007/s10858-013-9741-y.

54. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol*. 1997;273(1):283-298. doi:10.1006/jmbi.1997.1284.

55. Güntert P. Automated NMR structure calculation with CYANA. *Methods Mol Biol*. 2004;278:353-378. doi:10.1385/1-59259-809-9:353.

56. Brunger AT, Adams PD, Clore GM, et al. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Cryst*. 1998;54:905-921. doi:10.1107/S0907444998003254.

57. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. *Proteins Struct Funct Genet*. 2007;66(4):778-795. doi:10.1002/prot.21165.

58. Laskowski R a., MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*. 1993;26(November):283-291. doi:10.1107/S0021889892009944.

59. Davis IW, Leaver-Fay A, Chen VB, et al. MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res*. 2007;35(SUPPL.2):375-383. doi:10.1093/nar/gkm216.

60. Vallurupalli P, Bouvignies G, Kay LE. Studying “invisible” excited protein

states in slow exchange with a major state conformation. *J Am Chem Soc*. 2012;134(19):8148-8161. doi:10.1021/ja3001419.

61. Hansen DF, Vallurupalli P, Kay LE. An improved 15N relaxation dispersion experiment for the measurement of millisecond time-scale dynamics in proteins. *J Phys Chem B*. 2008;112(19):5898-5904. doi:10.1021/jp074793o.

62. Yip GNB, Zuiderweg ERP. A phase cycle scheme that significantly suppresses offset-dependent artifacts in the R2-CPMG 15N relaxation experiment. *J Magn Reson*. 2004;171(1):25-36. doi:10.1016/j.jmr.2004.06.021.

63. Morin S, Linnet TE, Lescanne M, et al. Relax: The analysis of biomolecular kinetics and thermodynamics using NMR relaxation dispersion data. *Bioinformatics*. 2014;30(15):2219-2220. doi:10.1093/bioinformatics/btu166.

64. Walker O, Varadan R, Fushman D. Efficient and accurate determination of the overall rotational diffusion tensor of a molecule from 15N relaxation data using computer program ROTDIF. *J Magn Reson*. 2004;168(2):336-345. doi:10.1016/j.jmr.2004.03.019.

65. Berlin K, Longhini A, Dayie TK, Fushman D. Deriving quantitative dynamics information for proteins and RNAs using ROTDIF with a graphical user interface. *J Biomol NMR*. 2013;57(4):333-352. doi:10.1007/s10858-013-9791-1.

66. Rovnyak D, Hoch JC, Stern a S, Wagner G. Resolution and sensitivity of high field nuclear magnetic resonance spectroscopy. *J Biomol NMR*. 2004;30(1):1-10. doi:10.1023/B:JNMR.0000042946.04002.19.

67. Brüschweiler R, Zhang F. Covariance nuclear magnetic resonance spectroscopy. *J Chem Phys*. 2004;120(2004):5253-5260. doi:10.1063/1.1647054.

68. Brüschweiler R. Theory of covariance nuclear magnetic resonance spectroscopy. *J Chem Phys*. 2004;121(2004):409-414. doi:10.1063/1.1755652.

69. Trbovic N, Smirnov S, Zhang F, Brüschweiler R. Covariance NMR spectroscopy by singular value decomposition. *J Magn Reson*. 2004;171:277-283. doi:10.1016/j.jmr.2004.08.007.

70. Zhang F, Brüschweiler R. Spectral deconvolution of chemical mixtures by covariance NMR. *ChemPhysChem*. 2004;5:794-796. doi:10.1002/cphc.200301073.

71. Zhang F, Brüschweiler R. Indirect covariance NMR spectroscopy. *J Am Chem Soc*. 2004;126(Figure 1):13180-13181. doi:10.1021/ja047241h.

72. Blinov K a., Larin NI, Kvasha MP, Moser A, Williams AJ, Martin GE. Analysis and elimination of artifacts in indirect covariance NMR spectra via unsymmetrical processing. *Magn Reson Chem*. 2005;43(12):999-1007.

doi:10.1002/mrc.1674.

73. Blinov K a., Larin NI, Williams AJ, Zell M, Martin GE. Long-range carbon-carbon connectivity via unsymmetrical indirect covariance processing of HSQC and HMBC NMR data. *Magn Reson Chem*. 2006;44(2):107-109. doi:10.1002/mrc.1766.

74. Kupče E, Freeman R. Hyperdimensional NMR spectroscopy. *J Am Chem Soc*. 2006;128(18):6020-6021. doi:10.1021/ja0609598.

75. Lescop E, Brutscher B. Hyperdimensional protein NMR spectroscopy in peptide-sequence space. *J Am Chem Soc*. 2007;129(39):11916-11917. doi:10.1021/ja0751577.

76. Benison G, Berkholz DS, Barbar E. Protein assignments without peak lists using higher-order spectra. *J Magn Reson*. 2007;189(2):173-181. doi:10.1016/j.jmr.2007.09.009.

77. Chen K, Delaglio F, Tjandra N. A practical implementation of cross-spectrum in protein backbone resonance assignment. *J Magn Reson*. 2010;203(2):208-212. doi:10.1016/j.jmr.2009.12.018.

78. Snyder DA, Brüschweiler R. Generalized indirect covariance NMR formalism for establishment of multidimensional spin correlations. *J Phys Chem A*. 2009;113:12898-12903. doi:10.1021/jp9070168.

79. Snyder DA, Ghosh A, Zhang F, Szyperski T, Brüschweiler R. Z-matrix formalism for quantitative noise assessment of covariance nuclear magnetic resonance spectra. *J Chem Phys*. 2008;129(10):1-9. doi:10.1063/1.2975206.

80. Snyder DA, Zhang F, Brüschweiler R. Covariance NMR in higher dimensions: Application to 4D NOESY spectroscopy of proteins. *J Biomol NMR*. 2007;39:165-175. doi:10.1007/s10858-007-9187-1.

81. Snyder DA, Xu Y, Yang D, Brüschweiler R. Resolution-enhanced 4D 15N/13C NOESY protein NMR spectroscopy by application of the covariance transform. *J Am Chem Soc*. 2007;129(46):14126-14127. doi:10.1021/ja075533n.

82. Nietlispach D, Ito Y, Laue ED. A novel approach for the sequential backbone assignment of larger proteins: Selective intra-HNCA and DQ-HNCA. *J Am Chem Soc*. 2002;124(2):11199-11207. doi:10.1021/ja025865m.

83. Permi P. Intraresidual HNCA: An experiment for correlating only intraresidual backbone resonances. *J Biomol NMR*. 2002:201-209.

84. Brutscher B. Intraresidue HNCA and COHNCA experiments for protein backbone resonance assignment. *J Magn Reson*. 2002;159:155-159. doi:10.1006/jmre.2002.2546.

85. Nietlispach D. A selective intra-HN (CA)CO experiment for the backbone

assignment of deuterated proteins. *J Biomol NMR*. 2004;28:131-136. doi:10.1023/B:JNMR.0000013829.17620.39.

86. Bodenhausen G. Direct determination of rate constants of slow dynamic processes by two-dimensional“ accordion” spectroscopy in nuclear magnetic resonance. *J Am Chem Soc*. 1982;3(21):1304-1309.

87. Gunther U, Ludwig C, Ruterjans H. NMRLAB-Advanced NMR data processing in matlab. *J Magn Reson*. 2000;145(2):201-208. doi:10.1006/jmre.2000.2071.

88. Barkhuijsen H. Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals. *J Magn Reson*. 1987;557:553-557.

89. Herman P, Lee JC. The Advantage of Global Fitting of Data Involving Complex Linked Reactions. In: Fenton AW, ed. *Allostery*. Vol 796. Methods in Molecular Biology. New York, NY: Springer New York; 2012:399-421. doi:10.1007/978-1-61779-334-9.

90. Harden BJ, Frueh DP. Molecular Cross-Talk between Nonribosomal Peptide Synthetase Carrier Proteins and Unstructured Linker Regions. *Angew Chemie Int Ed [submitted]*. 2016.

91. Strieker M, Tanović A, Marahiel MA. Nonribosomal peptide synthetases: Structures and dynamics. *Curr Opin Struct Biol*. 2010;20(2):234-240. doi:10.1016/j.sbi.2010.01.009.

92. Drake EJ, Miller BR, Shi C, et al. Structures of two distinct conformations of holo-non-ribosomal peptide synthetases. *Nature*. 2016;529(7585):235-238. doi:10.1038/nature16163.

93. Reimer JM, Aloise MN, Harrison PM, Schmeing TM. Synthetic cycle of the initiation module of a formylating nonribosomal peptide synthetase. *Nature*. 2016;529(7585):239-242. doi:10.1038/nature16503.

94. Goodrich AC, Harden BJ, Frueh DP. Solution Structure of a Nonribosomal Peptide Synthetase Carrier Protein Loaded with Its Substrate Reveals Transient, Well-Defined Contacts. *J Am Chem Soc*. 2015;137(37):12100-12109. doi:10.1021/jacs.5b07772.

95. Miller BR, Gulick AM. Structural Biology of Nonribosomal Peptide Synthetases. *Methods Mol Biol*. 2016;1401(May):3-29. doi:10.1007/978-1-4939-3375-4_1.

96. Lipari G, Szabo A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *J Am Chem Soc*. 1982;2(1).

97. Lipari G, Szabo A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J Am Chem Soc*. 1982;104(17):4546-4559.

doi:10.1021/ja00381a009.

98. Miller BR, Sundlov J a, Drake EJ, Makin T a, Gulick AM. Analysis of the linker region joining the adenylation and carrier protein domains of the modular nonribosomal peptide synthetases. *Proteins*. 2014;(May):1-12. doi:10.1002/prot.24635.

99. Williamson MP. Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc*. 2013;73:1-16. doi:10.1016/j.pnmrs.2013.02.001.

100. Terada T, Ito Y, Shirouzu M, et al. Nuclear magnetic resonance and molecular dynamics studies on the interactions of the Ras-binding domain of Raf-1 with wild-type and mutant Ras proteins. *J Mol Biol*. 1999;286(1):219-232. doi:10.1006/jmbi.1998.2472.

101. Tamiola K, Acar B, Mulder FAA. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J Am Chem Soc*. 2010;132(51):18000-18003. doi:10.1021/ja105656t.

102. Palmer AG. Chemical exchange in biomacromolecules: past, present, and future. *J Magn Reson*. 2014;241:3-17. doi:10.1016/j.jmr.2014.01.008.

103. Tollinger M, Skrynnikov NR, Mulder F a, Forman-Kay JD, Kay LE. Slow dynamics in folded and unfolded states of an SH3 domain. *J Am Chem Soc*. 2001;123(46):11341-11352.

104. Bae SH, Dyson HJ, Wright PE. Prediction of the rotational tumbling time for proteins with disordered segments. *J Am Chem Soc*. 2009;131(19):6814-6821. doi:10.1021/ja809687r.

105. Sundlov JA, Shi C, Wilson DJ, Aldrich CC, Gulick AM. Structural and functional investigation of the intermolecular interaction between NRPS adenylation and carrier protein domains. *Chem Biol*. 2012;19(2):188-198. doi:10.1016/j.chembiol.2011.11.013.

106. Weber T, Marahiel MA. Exploring the domain structure of modular nonribosomal peptide synthetases. *Structure*. 2001;9(1):3-9. doi:10.1016/S0969-2126(00)00560-8.

107. Carver JP, Biophysics M. A General Two-Site Solution for the Chemical Exchange Produced Dependence of T2 Upon the Carr-Purcell Pulse Separation. *J Magn Reson*. 1972;105:89-105.

8 Curriculum vitae

Bradley James Harden
Born 1988/02/22, Palm Beach Gardens, FL. USA

Educational History:

Ph.D.	2017	Biomedical Engineering Mentor: Dominique Frueh, PhD	Johns Hopkins University
B.S.	2010	Electrical Engineering	University of Florida

External Funding:

2010 – 2012	T32 Training Grant	NIH

Awards & Honors:

2015	Innovation Fellow	Thread
2013	Travel Award	Johns Hopkins

Publications:

Harden BJ, Frueh DP (2016) Molecular Cross-Talk between Nonribosomal Peptide Synthetase Carrier Proteins and Unstructured Linker Regions. ChemBioChem [in revision].

Harden BJ, Frueh DP (2016) Covariance NMR processing and analysis for protein assignment. Methods Mol Biol [in review].

Zeng H, Xu J, Yadav NN, McMahon MT, **Harden BJ**, Frueh DP, & van Zijl PCM. (2016). 15N Heteronuclear Chemical Exchange Saturation Transfer MRI. Journal of the American Chemical Society, 138(35), 11136–11139.

Goodrich AC, **Harden BJ**, & Frueh DP (2015). Solution Structure of a Nonribosomal Peptide Synthetase Carrier Protein Loaded with Its Substrate Reveals Transient, Well-Defined Contacts. Journal of the American Chemical Society, 137(37), 12100–12109.

Harden BJ, Mishra SH, & Frueh DP (2015). Effortless assignment with 4D covariance sequential correlation maps. Journal of Magnetic Resonance, 260, 83–88.

Mishra SH, **Harden BJ**, & Frueh DP (2014). A 3D time-shared NOESY experiment designed to provide optimal resolution for accurate assignment of NMR distance restraints in large proteins. Journal of Biomolecular NMR, 60(4), 265–274.

Harden BJ, Nichols SR, & Frueh DP (2014). Facilitated Assignment of Large Protein NMR Signals with Covariance Sequential Spectra Using Spectral Derivatives. Journal of the American Chemical Society, 136(38), 13106–13109.

Harden BJ, & Frueh DP (2014). SARA: a software environment for the analysis of relaxation data acquired with accordion spectroscopy. Journal of Biomolecular NMR, 58(2), 83–99.

Huang H, Tan BZ, Shen Y, Tao J, Jiang F, Sung YY, Ng CK, Raida M, Köhr G, Higuchi M, Fatemi-Shariatpanahi H, **Harden BJ**, Yue DT, Soong TW (2012). RNA editing of the IQ domain in Ca(v)1.3 channels modulates their Ca2+-dependent inactivation. Neuron, 73(2), 304–16.

Posters:

Harden BJ, Nichols SR, Mishra SH, Frueh DP (2016) Assigning backbone and sidechain resonances with 4D covariance spectra. Experimental NMR Conference

Harden BJ, Frueh DP, (2015) SARA – Acquire and analyze NMR relaxation data in one sixth of the time. Biophysical Society

Harden BJ, Frueh DP, (2013) SARA – Acquire and analyze NMR relaxation data in one sixth of the time. Eastern Analytic Symposium

Service & Leadership:

2015 – 2017 Thread Manager
Recruited & managed volunteers working with 8 students

2015 Thread Consultant
Selected from 900 volunteers to consult on organizational strategy

2014 Teaching Assistant
Models & Simulations

2013 Teaching Assistant
Thermodynamics & Statistical Mechanics

2013 – 2015 Thread Team Leader
Lead team to help student in & out of the classroom

2012 Thread Volunteer
Helped underachieving student with academics & life skills